

NORTHWESTERN UNIVERSITY

Computational Conceptual Change: An Explanation-Based Approach

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Scott E. Friedman

EVANSTON, ILLINOIS

June 2012

Report Documentation Page		Form Approved OMB No. 0704-0188
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.		
1. REPORT DATE <b>JUN 2012</b>	2. REPORT TYPE	3. DATES COVERED <b>00-00-2012 to 00-00-2012</b>
4. TITLE AND SUBTITLE <b>Computational Conceptual Change: An Explanation-Based Approach</b>		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER	
	5e. TASK NUMBER	
	5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>Northwestern University, 2133 Sheridan Road, Evanston, IL, 60208</b>		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>		
13. SUPPLEMENTARY NOTES		
14. ABSTRACT <p><b>The process of conceptual change ? whereby new knowledge is adopted in the presence of prior conflicting knowledge ? is pervasive in human cognitive development, and contributes to our cognitive flexibility. At present, Artificial Intelligence systems lack the flexibility of human conceptual change. This is due in part to challenges in knowledge representation, belief revision abduction, and induction. In addition, there are disagreements in the cognitive science community regarding how people represent, use, and revise their mental models of the world. This work describes a cognitive model of conceptual change. The claims are that (1) qualitative models provide a consistent computational account of human mental models, (2) our psychologically plausible model of analogical generalization can learn these models from examples, and (3) conceptual change can be modeled by iteratively constructing explanations and using meta-level reasoning to select among competing explanations and revise domain knowledge. The claims are supported by a computational model of conceptual change, an implementation of our model on a cognitive architecture, and four simulations. We simulate conceptual change in the domains of astronomy, biology, and force dynamics where examples of psychological conceptual change have been empirically documented. Aside from demonstrating domain generality, the simulations provide evidence for the claims of the thesis. Our simulation that learns mental models from observation induces qualitative models of movement, pushing, and blocking from observations and performs similar to students in problem-solving. Our simulation that creates and revises explanations about the changing of the seasons shows that our system can assemble and transform mental models like students. Our simulation of textbook knowledge acquisition shows that our system can incrementally repair incorrect knowledge like students using self-explanation. Finally, our simulation of learning and revising a force-like concept from observations shows that our system can use heuristics and abduction to revise quantities in a similar manner as people. The performance of the simulations provides evidence of (1) the accuracy of the cognitive model and (2) the adaptability of the underlying cognitive systems that are capable of conceptual change.</b></p>		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT <b>Same as Report (SAR)</b>	18. NUMBER OF PAGES <b>302</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

© Copyright by Scott E. Friedman 2012  
All Rights Reserved

## ABSTRACT

### Computational Conceptual Change: An Explanation-Based Approach

Scott Friedman

The process of conceptual change – whereby new knowledge is adopted in the presence of prior, conflicting knowledge – is pervasive in human cognitive development, and contributes to our cognitive flexibility. At present, Artificial Intelligence systems lack the flexibility of human conceptual change. This is due in part to challenges in knowledge representation, belief revision, abduction, and induction. In addition, there are disagreements in the cognitive science community regarding how people represent, use, and revise their mental models of the world.

This work describes a cognitive model of conceptual change. The claims are that (1) qualitative models provide a consistent computational account of human mental models, (2) our psychologically plausible model of analogical generalization can learn these models from examples, and (3) conceptual change can be modeled by iteratively constructing explanations and using meta-level reasoning to select among competing explanations and revise domain knowledge. The claims are supported by a computational model of conceptual change, an implementation of our model on a cognitive architecture, and four simulations.

We simulate conceptual change in the domains of astronomy, biology, and force dynamics, where examples of psychological conceptual change have been empirically documented. Aside from demonstrating domain generality, the simulations provide evidence for the claims of the thesis. Our simulation that learns mental models from observation induces qualitative models of movement, pushing, and blocking from observations and performs similar to students in problem-solving. Our simulation that creates and revises explanations about the changing of the

seasons shows that our system can assemble and transform mental models like students. Our simulation of textbook knowledge acquisition shows that our system can incrementally repair incorrect knowledge like students using self-explanation. Finally, our simulation of learning and revising a force-like concept from observations shows that our system can use heuristics and abduction to revise quantities in a similar manner as people. The performance of the simulations provides evidence of (1) the accuracy of the cognitive model and (2) the adaptability of the underlying cognitive systems that are capable of conceptual change.

## Acknowledgments

The five years I've spent pursuing my Ph.D. have been some of the most rewarding years in my life. Although Artificial Intelligence has interested me from a young age, this endeavor would not have been as rewarding without the support of my family, my friends, my colleagues, and friendly coffee establishments such as Urban Orchard. I thank Northwestern University, the NU Cognitive Science program, and the Office of Naval Research for funding my research.

I am grateful to my advisor Ken Forbus for his guidance and his company. I imagine that all graduate students require some coaching to harness their potential, so if I've grown as a scientist or as a writer, Ken deserves his share of the responsibility. Ken's passion and optimism are contagious, and provided me the boost I needed amidst publication deadlines and dissertation revisions. I believe one can be a good scientist *without* having this passion in one's work, but Ken is an example of how fulfilled and productive one can be when passion and expertise align. I hope to remember this for whatever lies ahead, in research or in life.

My thesis committee was a pleasure to work with. I thank Dedre Gentner for useful discussions about analogical reasoning and cognitive modeling. Dedre helped me realize the interdisciplinary potential of Artificial Intelligence. I thank Bryan Pardo for helping me sustain a sense of humor throughout my graduate career, Christopher Riesbeck for helping me improve my argumentation, and Bruce Sherin for discussing conceptual change and for collaborating during my cognitive science advanced research fellowship.

It pains me to imagine parting ways with my fellow researchers in the Qualitative Reasoning Group. We have helped each other in times of distress and uncertainty, shared high-brow tastes, and exchanged more low-brow humor than I care to admit. Thanks to Klenk, Dehghani, Tomai,

Lockwood, and Lovett for their collaboration and coaching, despite having their own agendas of publication and graduation. Thanks to Jon Wetzel for being my wingman in the Ph.D. program, and to Thomas Hinrichs for keeping the group's wheels from falling off, while still managing to advise confused students and shout new and creative obscenities at a computer. Thanks to Matt McLure for daring to ask questions (since I think I benefited equally from trying to answer them) and to Jason Taylor for filling the lab with camaraderie and empty cans of Red Bull. Maria Chang has been a great friend to me – we shared an office, a whiteboard, weekend updates, recipes, bourbon whiskey, music, gossip, and dirty silverware. By my decree, Maria is not allowed to share an office with anybody else after my departure, since I don't want to miss the laughs and the stories. I'll miss all of my lab-mates very much.

My parents and my in-laws have taken a genuine – or, at least, convincing – philosophical interest in my research, and I deeply appreciated their cheerleading and reassurance while I wrote my dissertation and hunted for jobs. There were months when my parents and my sister heard my answering machine more than they should have, and I am grateful for their patience and understanding.

My wife Sara has been the ideal partner in crime for the last five years. She drags me outdoors to take walks in the sun, she makes parenting a delight, she dealt with the geographical uncertainty of my career, and she is genuinely very proud of me. Despite the time we spent together, finding time to do research was never a problem; Sara's 80-hour work-weeks during residency permitted me plenty of time to attend to my research, and our daughter Ada's 5am crying fits during our month of sleep-training bought me some extra crunch-time near publication deadlines. I can't imagine life without Sara and Ada. Ada: if you're reading this, come find me – I'll drop whatever I'm doing and we'll go get ice cream. Mom can come too.



This dissertation is dedicated to  
two friends whose curiosity will always inspire me:  
John Krettek, III (1981-2009) and Ada Marie Friedman (2010-).

## Contents

ABSTRACT.....	3
Acknowledgments.....	5
List of Figures .....	14
Chapter 1: Introduction .....	18
1.1 Claims .....	24
1.2 Psychological assumptions of our model of conceptual change.....	31
1.2.1 Assumptions about knowledge representation .....	33
1.2.2 Assumptions about memory and knowledge organization.....	36
1.2.3 Assumptions about learning .....	39
Chapter 2: Other theories of conceptual change .....	45
2.1 Four theories of conceptual change .....	46
2.1.1 Carey’s theory-theory of conceptual development.....	46
2.1.2 Vosniadou’s framework theory .....	48
2.1.3 Chi’s categorical shift.....	50
2.1.4 diSessa’s knowledge in pieces.....	53
2.2 Divergence and disagreement.....	55
2.2.1 What counts as conceptual change .....	55
2.2.2 Revision versus addition.....	56
2.2.3 The effect of explaining on the process of change .....	64

2.2.4 The source of coherence .....	67
2.3 The path forward .....	70
Chapter 3: Background .....	71
3.1 Ontologies .....	71
3.2 Qualitative reasoning .....	72
3.2.1 Qualitative Process Theory .....	74
3.2.2 Compositional modeling .....	75
3.3 Abductive reasoning .....	80
3.4 Analogical processing .....	81
3.4.1 The Structure-Mapping Engine .....	82
3.4.2 MAC/FAC .....	83
3.4.3 SAGE .....	84
3.5 Truth Maintenance Systems .....	86
3.6 Microtheory contextualization .....	88
3.7 Metareasoning .....	90
3.8 CogSketch .....	91
3.8.1 Psychological assumptions about using comic graphs .....	92
Chapter 4: A Computational Model of Explanation-Based Conceptual Change .....	96
4.1 Two micro-examples of conceptual change .....	96
4.2 Contextualizing knowledge for conceptual change .....	98

	10
4.3 An explanation-based network for conceptual change .....	104
4.4 Constructing explanations.....	107
4.4.1 Psychological assumptions of explanation construction .....	124
4.4.2 Explanation competition.....	125
4.5 Explanation retrieval and reuse.....	126
4.6 Finding the preferred explanation.....	129
4.6.1 Rule-based epistemic preferences.....	130
4.6.2 Cost functions .....	137
4.7 Retrospective explanation.....	147
Chapter 5: Learning intuitive mental models of motion from observation .....	150
5.1 Using multimodal training data .....	152
5.2 Creating generalizations of Pushing, Moving, and Blocking with SAGE.....	156
5.3 Converting SAGE generalizations to qualitative models .....	157
5.4 Comparing the system's models of motion to students' mental models .....	161
5.5 Discussion.....	166
Chapter 6: Revising mechanism-based models of the seasons .....	168
6.1 How commonsense explanations (and seasons) change.....	169
6.2 Simulating how students construct and revise explanations.....	173
6.2.1 Explaining Chicago's seasons .....	174
6.2.2 Explaining Australia's seasons .....	182

6.2.3 Comparing the system's explanations to student explanations .....	182
6.2.4 Accommodating new, credible information .....	183
6.3 Simulation results .....	189
6.4 Discussion .....	192
Chapter 7: Mental model transformation from textbook information .....	195
7.1 Self-explaining improves student accommodation of textbook material .....	197
7.2 Simulating the self-explanation effect .....	200
7.2.1 Explanandums: situations that require an explanation .....	202
7.2.2 Constructing explanations to generate the pre-instructional model .....	203
7.2.3 Determining the simulation's circulatory model .....	206
7.2.4 Integrating textbook information .....	207
7.2.5 Assuming model participants .....	215
7.3 Simulation results .....	217
7.4 Discussion .....	219
Chapter 8: Revising a category of force when explanations fail .....	222
8.1 Assessing the changing meaning of force in students .....	224
8.1.1 Replicating the force questionnaire and approximating students' observations .....	227
8.2 Learning by explaining new observations .....	228
8.2.1 Declarative heuristics for failure-based revision .....	229
8.2.2 Choosing among applicable heuristics .....	233

8.2.3 Revising conceptual quantities .....	233
8.2.4 Inter-scenario analysis .....	237
8.2.5 Retrospective explanation propagates revisions .....	242
8.3 Simulation results .....	243
8.4 Discussion .....	245
Chapter 9: Conclusion .....	249
9.1 Revisiting the claims .....	249
9.2 Related work in AI .....	252
9.3 Comparison to other theories of conceptual change .....	258
9.3.1 Knowledge in pieces .....	258
9.3.2 Carey's theory .....	259
9.3.3 Chi's categorical shift .....	260
9.3.4 Vosniadou's framework theory .....	261
9.3.5 Novel aspects of our model as a theory of conceptual change .....	262
9.4 Future work and limitations .....	264
9.4.1 Simulating over larger timescales .....	265
9.4.2 Improving explanation construction .....	267
9.4.3 Improving explanation evaluation .....	268
9.4.4 Other types of agency .....	269
9.4.5 Taking analogy further .....	270

9.4.6 Accruing domain knowledge .....	271
9.4.7 Storing explanations .....	272
9.4.8 Clustering explanandums.....	273
9.4.9 Proactivity .....	274
9.4.10 Applying the model of conceptual change .....	274
References.....	276
Appendix.....	293
Definitions .....	293
Transcript of an interview about the seasons from Chapter 6 .....	296
Rules for detecting contradictions .....	300
Sentences from a textbook passage about the circulatory system .....	301

## List of Figures

<b>Figure 1.</b> Correspondences between psychological and artificial entities in this dissertation. ....	24
<b>Figure 2.</b> Two intuitive accounts of the human circulatory system. They share propositional beliefs and mental model parts, but some propositional beliefs in $D_a$ are inconsistent with those in $D_b$ . ....	27
<b>Figure 3.</b> High-level psychological assumptions of our cognitive model, organized by type. Each is labeled where supported by (s) or unsupported by (u) the literature. ....	32
<b>Figure 4:</b> Five possible accounts of how category information is revised. Black and white nodes represent categories and phenomena, respectively. Arrows indicate “is understood in terms of.” Dotted zones indicate contexts. (a) Initial state with category Force[1]. (b-f) Possible resultant states after incorporating Force[2]. ....	58
<b>Figure 5:</b> A small portion of the OpenCyc ontology. An arrow $a \rightarrow b$ indicates $(\text{genls } a \ b)$ . ....	72
<b>Figure 6:</b> ContainedFluid (above) and FluidFlow (below) model fragments used in the simulation in Chapter 7. English interpretations for the model fragments included at right. ....	75
<b>Figure 7:</b> A TMS containing assumptions (squares), justified beliefs (ovals), justifications (triangles), and a contradiction $\perp$ node (courtesy Forbus & de Kleer, 1993) ....	86
<b>Figure 8:</b> Meta-level control and monitoring (Cox & Raja, 2007) ....	90
<b>Figure 9:</b> A sketched comic graph stimulus. ....	91
<b>Figure 10:</b> A sketch with two subsketches, redrawn from diSessa et al. (2004). ....	93
<b>Figure 11:</b> A small portion of justification structure generated from model formulation in the circulatory system micro-example. The justification (triangle) at left is the logical instantiation of model fragment instance $mfi0$ based on the constraints of ContainedFluid (see Figure 6 for ContainedFluid definition) and the justification at right is the logical activation of $mfi0$ . ....	100
<b>Figure 12:</b> The relationship between microtheories (MTs) in our computational model. Solid arrows represent “inherit all information from” (i.e., child-of), and dotted arrows represent “contains some information from.” ....	102
<b>Figure 13:</b> A portion of an explanation-based network. (a) Single explanation $x_0$ for an explanandum $naiveH2B$ (rightmost nodes). (b) After new knowledge is added, preferences are computed for new knowledge ( $<_c$ ), new model fragment instances ( $<_{mfi}$ ), and for the new explanation $x_1$ ( $<_{xp}$ ). ....	104



<b>Figure 14:</b> ContainedFluid (above) and FluidFlow (below) model fragments used in the simulation in Chapter 7. English interpretations of each model fragment (at right). .....	108
<b>Figure 15:</b> Pseudo-code for front-end procedures that trigger abductive model formulation. ..	111
<b>Figure 16:</b> A portion of explanation-based network. (a) Before an explanation has been constructed for naiveH2B. (b) After an explanation $x_0$ has been constructed for naiveH2B via abductive model formulation. ....	113
<b>Figure 17:</b> Pseudo-code for abductive model instantiation.....	117
<b>Figure 18:</b> The network after two explanations have been constructed via abductive model formulation: $x_0$ explains naiveH2B, and $x_1$ explains naiveH2B and leftH2B. ....	120
<b>Figure 19:</b> A graph of the relationships between model fragments and other collections in the circulatory system example.....	121
<b>Figure 20:</b> Pseudo-code for best explanation retrieval algorithms, which use MAC/FAC to find explanations that are relevant for a given explanandum or case. ....	128
<b>Figure 21:</b> Pseudo-code for computing an explanation's cost and a belief's cost using a cost function. Note that the cost of any explanation that is presently adopted (i.e., an explanandum is mapped to it in $\mathbb{E}$ ) is zero. ....	141
<b>Figure 22:</b> A sequence of events from the simulation in Chapter 6 that produces several contradictions between best explanations and credible domain knowledge.....	143
<b>Figure 23:</b> Algorithm for restructuring knowledge based on the presence of a high-cost artifact. ....	144
<b>Figure 24:</b> Model fragment ArterialFlow is preferred over FluidFlow due to greater specificity, but leftH2B has not yet been explained using the preferred knowledge.....	148
<b>Figure 25:</b> Topology of the Chapter 5 simulation.....	151
<b>Figure 26:</b> A sketched comic graph stimulus.....	153
<b>Figure 27:</b> The three SAGE generalization contexts after using SAGE to generalize temporally-encoded microtheories about pushing, moving, and blocking.....	156
<b>Figure 28:</b> Given a statement $s$ and its temporal relationship to an event $e$ , how to calculate the causal role(s) of $s$ in a qualitative model of $e$ . ....	158
<b>Figure 29:</b> One of the qualitative models learned by the simulation that causally relates pushing and movement. Summaries of constraints and consequences shown at right.....	159
<b>Figure 30:</b> The sketch for the problem-solving task from Brown (1994). ....	163

**Figure 31:** Problem from the Force Concept Inventory, and student/simulation responses (left). Sketch of the same problem using CogSketch (right). ..... 165

**Figure 32:** Two diagrams explaining seasonal change, courtesy of Sherin et al. (2012). (a) Sketch from a novice student, explaining that the earth is closer to the sun in the summer than in the winter. (b) Scientific explanation involving tilt and insolation. .... 170

**Figure 33:** AstronomicalHeating (top) and Approaching-PeriodicPath (bottom) model fragments used in the simulation. English interpretations of both model fragments included at right. .... 172

**Figure 34:** Pseudo-code for constructing explanations about ordinal relations and quantity changes, from Chapter 4. .... 176

**Figure 35:** Network plotting explanations  $x_0$  and  $x_1$  that explain seasonal change in Australia ( $x_0$ ) and Chicago ( $x_1$ ) using a near-far model of the seasons. .... 179

**Figure 36:** An influence diagram of the near-far explanation of both Chicago's (Chi) and Australia's (Aus) seasons. Nodes are quantities and edges describe positive and negative direct influences (i+, i-) and indirect influences (q+, q-). Bracketed ranges quantify process activity. .... 183

**Figure 37:** An influence diagram of the facing explanation of both Chicago's (Chi) and Australia's (Aus) seasons. .... 188

**Figure 38:** Influence graphs for additional explanations produced by the simulation. (a) The tilt of the axis increases and decreases each hemisphere's distance to the sun. (b) A simplified correct explanation of the seasons. .... 190

**Figure 39:** Student models of the human circulatory system from Chi et al. (1994a). .... 198

**Figure 40:** Transitions between pretest and posttest models for control and prompted groups in Chi et al. (1994a). Numbers indicate the number of students who made the given transition. See Figure 39 for an illustration of each mental model. .... 199

**Figure 41:** ContainedFluid (above) and FluidFlow (below) model fragments used in the simulation. English interpretations of each model fragment (at right). .... 201

**Figure 42:** A portion of explanation-based network. (a) Before an explanation has been constructed for naiveH2B. (b) After an explanation  $x_0$  has been constructed for naiveH2B via abductive model formulation. .... 204

**Figure 43:** Influence graphs generated by the system to describe the relative concentrations, infusion, and consumption of Oxygen. Left: using "Double loop (1)" model. Right: using "Double loop (2)" model. .... 206

**Figure 44:** Portion of explanation-based network. (a): After explaining blood flow from heart to body (naiveH2B). (b): After explaining blood flow from the left-heart to the body (leftH2B),

with preferences across concepts ( $<_c$ ), model fragment instances ( $<_{mfi}$ ), and explanations ( $<_{xp}$ ). .....	215
<b>Figure 45:</b> Circulatory model transitions for all simulation trials.....	218
<b>Figure 46:</b> Occurrences of meaning of force, by grade.....	225
<b>Figure 47:</b> At left: a four-frame comic graph used as training data. ....	226
<b>Figure 48:</b> Left: a heuristic <code>createDecreaseProcess</code> that automatically creates a new model fragment to explain a quantity decreasing. Right: Process model of leftward movement $m_1$ automatically created with this heuristic. ....	229
<b>Figure 49:</b> Left: a heuristic <code>addHiddenQtyCond</code> that revises process models by adding a hidden (conceptual) quantity.....	231
<b>Figure 50</b> (a) Model fragment $m_1$ (Figure 48, right) explains the ball moving, but not the ball stopping. (b) After revising $m_1$ as $m_2$ (Figure 49, right), $m_2$ explains both phenomena, and preferences are computed. ....	231
<b>Figure 51:</b> Left: a heuristic <code>vectorizeQty</code> that transforms a scalar conceptual quantity into a vector quantity and revises the according model fragment to take a direction.....	234
<b>Figure 52:</b> Comic graph scenarios A and B are sufficiently similar for inter-scenario analysis. .....	238
<b>Figure 53:</b> Selected analogical correspondences between Scenarios A and B (Figure 52). ....	239
<b>Figure 54:</b> Changes in the simulation's meaning of force, using Ioannides and Vosniadou's (2002) student meanings of force. ....	244
<b>Figure 55:</b> Using SAGE to cluster explanandums so that one explanation can justify multiple observations that are close analogs of one another. ....	272

## Chapter 1: Introduction

“We are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction.”

- Otto Neurath (in Quine, 1960)

Neurath’s analogy between rebuilding a ship at sea and lifelong learning communicates several important insights. Like sailors reconstructing their ship, we can repair our intuitive knowledge to become more scientifically correct. We are constrained by the need for support: as beams on the ship require the support of adjacent beams, so does our understanding of observations rely on the support of explanations. Consequently, the transformations of the ship and our knowledge involve the revision of components and the transition of support. In cognitive science, this transformation process is known as *conceptual change*. Following diSessa’s (2006) characterization, conceptual change is the process of building new ideas in the context of existing, conflicting ideas. This is differentiated from skill learning (since skills involve procedural knowledge) and from the *tabula rasa* acquisition of knowledge (hence the emphasis on “change”). This also does not include filling gaps in incomplete knowledge (Chi, 2008) or enriching (i.e., adding detail to) existing knowledge (Carey, 1991). We provide examples of conceptual change to help illustrate.

One well-documented example of conceptual change is the changing concept of force in children (Ioannides and Vosniadou, 2002; diSessa et al., 2004). When students enter the classroom, they have intuitive concepts of force learned from experience and interaction. One intuitive theory is that forces act on objects to keep them translating or rotating, and then gradually die off – similar to the medieval concept of *impetus* (McCloskey, 1983). While scientifically incorrect, this concept of force is still productive for understanding and manipulating the world, which is probably why it is so resilient to change. Through education, students revise these intuitive concepts incrementally, although often unsuccessfully. Even after learning scientifically correct *quantitative* aspects of force such as  $F = m \cdot a$ , students often operate with the same incorrect *qualitative* theories when labeling forces and drawing projectile trajectories (Clement, 1985; Hestenes et al., 1992).

Revising the concept of force involves revising the *specification* (diSessa et al., 2004) of the category. The specification includes the conditions under which a force exists, the consequences of a force's existence, how forces are combined, and the relationship of a force to other quantities (e.g., mass, velocity, acceleration).<sup>1</sup> For example, there is evidence that novices frequently conceive of force as a substance-like quantity (Reiner et al., 2000) that can be acquired, possessed, transferred, and subsequently lost by physical objects. Changing force from this intuitive, substance-like specification to a Newtonian specification requires changing the conditions and consequences of a force's existence, the model of how forces combine, and the relationship of force to acceleration and mass. We refer to this type of conceptual change as *category revision*, and we discuss this further in Chapter 8. An example of category revision is differentiating heat and temperature. This has been characterized in the history of science (Wiser and Carey, 1983) as well as within individual students (Wiser and Amin, 2001): the words “heat”

---

<sup>1</sup> Reif (1985) refers to this as the *ancillary knowledge* of a quantity.

and “temperature” are at first used interchangeably, and then this over-general concept is revised into two specific concepts, resulting in an intensive concept of temperature and an extensive concept of heat.

There is considerable disagreement among cognitive scientists on how this type of conceptual change occurs: do categories actually get directly shifted (e.g., Chi, 2008)? Are they added as additional categories that coexist alongside the prior category (e.g., diSessa and Sherin, 1998)? Do the new and the old categories coexist, but in different conceptual systems (e.g., Carey, 2009)?<sup>2</sup> If new and old categories coexist somehow, people seem to understand that they are mutually incoherent, perhaps due to belief-level refutation (Chi, 2008) or incompatibility between the vocabularies (Carey, 2009). Regardless of whether and how information coexists, any cognitive model of conceptual change must explain how people come to use a new conceptual vocabulary (e.g., Newtonian force) in place of an old vocabulary (e.g., impetus-like force).

The second type of conceptual change we simulate is *mental model transformation* (Chi, 2008). This involves revising the causal knowledge about physical systems in our long-term memory, which are often referred to as *mental models* (Gentner & Stevens, 1983). Suppose a student has the common misconception that blood flows in a single loop in the human circulatory system: from the heart to the rest of the body, and then back again, to be oxygenated by the heart (Chi et al., 1994a). Revising this mental model of the circulatory system to include a second loop – from the heart to the lungs for oxygenation, and then back – involves a transformation of this knowledge. This is not merely filling a gap in incomplete knowledge, since the old and new models of the circulatory system make conflicting predictions. This type of conceptual change has also been characterized in the domains of biology (Carey, 1985; Keil, 1994; Inagaki &

---

<sup>2</sup> Chapter 2 discusses this and other points of disagreement and divergence in theories of conceptual change.

Hatano, 2002), the shape of the earth (Vosniadou and Brewer, 1992), the changing of the seasons (Atwood & Atwood, 1996; Sherin et al., 2012), and others. Both types of conceptual change – category revision and mental model transformation – are the ubiquitous results of our attempts to reconcile new observations and instructions into our existing belief system.

Conceptual change is pervasive in our cognitive development and education, and contributes to the flexibility of human thought over time. The same is not true for Artificial Intelligence (AI) systems; at present, AI systems are *brittle* (McCarthy, 2007) in that they often malfunction when faced with new types of tasks and unexpected observations. Many researchers in the field believe this can be fixed by making the central cognitive architecture of AI systems more flexible and adaptable (e.g. Nilsson, 2005; Cassimatis, 2006). We believe that conceptual change is an important consideration for building more adaptable AI.

Modeling conceptual change will have a number of practical applications. For example, scientific discovery systems would benefit from having more flexible representations, whether using machines as collaborators (e.g., Langley, 2000), as automated scientists (e.g., Ross, 2009; Langley, 1983), or as mathematicians (e.g., Lenat & Brown, 1984). Intelligent tutoring systems will benefit similarly – if a tutoring system can model a student’s intuitive knowledge<sup>3</sup> and model the process of conceptual change, it can help guide the student through difficult learning. Finally, conceptual change will affect how we interact with intelligent agents. As eloquently put by Lombrozo (2006), explanations are the currency with which we exchange beliefs. Conceptual change – and explanation construction, which is part of our conceptual change model – will help AI systems exchange the same explanatory “currency” as people. Specifically, this will help an AI system (1) construct explanations that are understandable by humans, (2) represent

---

<sup>3</sup> See Anderson & Gluck (2001) for how one type of tutoring system models students’ procedural mathematics knowledge. Procedural knowledge is not included in our model of conceptual change.

explanations provided by humans and other resources (e.g., textbooks), and (3) revise beliefs and explanations as humans do, for more intuitive interaction.

Given these benefits to human-level AI research and applied AI systems, why not provide these systems with a computational model of human conceptual change? Unfortunately, such a computational model does not yet exist. I believe this is due to two general obstacles: (1) the complexity of human conceptual change and (2) disagreements in the cognitive science community how conceptual change occurs. Human conceptual change is complex in that it involves constructing explanations (Chi et al., 1994a), revising beliefs and explanations (Sherin et al., 2012; Vosniadou & Brewer, 1992), analogy (Gentner et al., 1997; Brown & Clement, 1989), and decision-making about new information (Chinn & Brewer, 1998). The major points of contention in the cognitive science literature involve the representation of conceptual knowledge (Forbus & Gentner, 1997; Nersessian, 2007), the organization of conceptual knowledge (diSessa et al., 2004; Ioannides & Vosniadou, 2002), and the mechanisms of change (Ohlsson, 2009; Chi and Brem, 2009; diSessa and Sherin, 1998). Fortunately, advances in cognitive science, both theoretical and empirical, have reached the point where modeling this complex phenomenon is now more feasible.

This dissertation presents and evaluates an integrated model of human conceptual change. The evaluation of our computational model and criteria for success rely upon its accuracy in explaining and predicting human learning and problem-solving. In each simulation, the system starts with similar knowledge as people, it is given similar stimuli for learning as people, and its knowledge is evaluated using similar problem-solving tasks as people. By comparing the system's problem-solving performance with those of students described in the literature, we can determine whether the system can learn along a humanlike trajectory of misconceptions and



scientific theories. We simulate different students by varying the system's starting knowledge and altering simulation parameters. Success is determined by the range of student trajectories our system can match using this strategy across simulation trials.

Our cognitive model makes a number of psychological assumptions concerning human perception, knowledge representation, reasoning, and learning. We hold these core assumptions constant across all four simulations, and describe them later in this chapter. In addition, each simulation makes task-specific assumptions. Some of these core and task-specific assumptions are needed to deal with current limitations of the conceptual model: for example, in some cases the model retains more information about a learning experience than is likely for humans. These interim assumptions provide explicit opportunities for extending this research. Half of the simulations use automatically generated training and testing data, and half use hand-coded data based on evidence from the literature. Both of types of data make assumptions about psychological knowledge encoding that are discussed below.

This dissertation is structured as follows. The rest of Chapter 1 is focused on the problem of conceptual change, the central theoretical claims of this dissertation, and the high-level psychological assumptions of this cognitive model. Chapter 2 discusses other theories of human conceptual change in the cognitive psychology literature. Chapter 3 reviews the AI theories and techniques used in our computational model. Chapter 4 presents the model of conceptual change and defines the terminology and algorithms used in the simulations. The model of conceptual change is a novel contribution of this dissertation, but it builds upon the existing AI technologies described in Chapter 3. Chapters 5-8 discuss four simulations: Learning intuitive mental models (Chapter 5); mental model transformation as explanation revision (Chapter 6); mental model transformation from a textbook passage (Chapter 7); and category revision for changing a

concept of force (Chapter 8). Chapter 9 revisits the claims, outlines some related work, and explores some objections, limitations, and opportunities for future work. The appendices contain additional algorithms and material for replication of the work described here.

	<b>Human</b>	<b>Our Model</b>
<b>Noun</b>	“human”	“system,” “simulation,” “AI”
<b>Adjective</b>	“psychological”	“artificial,” “computational”
<b>Models</b>	“mental model”	“compositional qualitative model”
<b>Model parts</b>	“mental model part”	“model fragment”
<b>Quantities</b>	← “quantity,” “quantity specification” →	
<b>Beliefs</b>	← “propositional belief” →	
<b>Explanations</b>	← “explanation” →	

**Figure 1. Correspondences between psychological and artificial entities in this dissertation.**

## 1.1 Claims

In this section we state the three principal claims of this dissertation and outline how these claims are supported. In discussing our claims and presenting our cognitive model, it is important to clarify when we are referring to people and when we are referring to AI systems. We include Figure 1 to prevent ambiguity. For the remainder of this dissertation, “human” and “psychological” will refer to humans, “AI” and “artificial” will refer to the computational model, and “agent” will refer to both. The first claim concerns how to represent human mental models in an AI system:

**Claim 1:** Compositional qualitative models provide a consistent computational account of human mental models.

By “consistent computational account” we mean that compositional qualitative models can consistently explain how people solve problems and construct explanations in multiple domains. Since Claim 1 is a knowledge representation claim, it can be tested by (1) observing how people construct explanations and solve problems with their mental models and (2) using compositional qualitative models to construct the same explanations and solve the same problems. Claim 1 is not a new idea – in fact, human mental models were one of the initial motivations for qualitative modeling in AI (Forbus & Gentner, 1997); however, we include this claim in the dissertation because we offer considerable novel evidence to support it (i.e., the simulations in Chapters 5-8) and the other claims rely upon it. We provide an overview of compositional qualitative models in AI in Chapter 3.

This dissertation includes a simulation of how people learn mental models from a sequence of observations, described in Chapter 5. With respect to Claim 1, this simulation uses qualitative models to simulate human mental models, but it also relies on an analogical learning algorithm called SAGE. SAGE is a psychologically plausible model of *analogical generalization* – that is, it abstracts the common relational structure across multiple cases. We discuss SAGE further in Chapter 3, but it is a component of the next claim.

**Claim 2:** Analogical generalization, as modeled by SAGE, is capable of inducing qualitative models that satisfy Claim 1.

Claim 2 is a novel claim, since AI systems have not previously induced qualitative models by these means. Claim 2 is supported by the simulation described in Chapter 5.

The third claim involves modeling the two types of conceptual change described above:

**Claim 3:** Human mental model transformation and category revision can both be modeled by iteratively (1) constructing explanations and (2) using meta-level reasoning to select among competing explanations and revise domain knowledge.

Claim 3 relies on the terms *explanation*, *meta-level*, and *domain knowledge*. We define these terms here with a simple example to clarify this claim. We define these same terms more precisely in Chapters 3 and 4. We intentionally avoid the word “theory” when referring to human knowledge, since this word has been used to describe (1) systematic science knowledge, (2) “intuitive theories” of novices, and (3) “domain theories” of model-based reasoning systems. We can thereby avoid conflating these distinct concepts.

We test Claim 3 by building a computational model and evaluating it according to the criteria put forth by Cassimatis, Bello, and Langley (2008): (1) the model’s ability to reason and learn as people do; (2) the breadth of situations in which it can do so, and (3) the parsimony of mechanisms it posits (i.e., using the same mechanisms across domains and tasks).

In this dissertation, *domain knowledge* is comprised of one or more of the following: propositional beliefs (i.e., a statement that evaluates to *true* or *false*), quantities (e.g., a specification of “force”), and mental model parts (see Figure 1 for modeling vocabulary and

Figure 2 for examples).<sup>4</sup> Consider the two sets of domain knowledge  $D_a$  and  $D_b$  about the human circulatory system in Figure 2 which are simplified accounts of student knowledge (Chi et al., 1994a).

	<b><math>D_a</math>: single loop</b>	<b><math>D_b</math>: double loop</b>
<b>Propositional beliefs</b>	Blood is a type of liquid The heart contains blood Arteries channel blood <i>from</i> the heart Veins channel blood <i>to</i> the heart ... The heart oxygenates blood <i>All</i> blood leaving the heart flows directly to the rest of the body <i>All</i> blood leaving the rest of the body flows directly to the heart	Blood is a type of liquid The heart contains blood Arteries channel blood <i>from</i> the heart Veins channel blood <i>to</i> the heart ... The lungs oxygenate blood <i>Some</i> blood leaving the heart flows directly to the rest of the body <i>All</i> blood leaving the rest of the body flows directly to the heart <i>Some</i> blood leaving the heart flows directly to the lungs <i>All</i> blood leaving the lungs flows directly to the heart
<b>Mental model parts</b>	Fluid-flow Infusing-compound-into-liquid Consuming-compound-from-liquid	Fluid-flow Infusing-compound-into-liquid Consuming-compound-from-liquid
<b>Quantity specs.</b>	<i>none</i>	<i>none</i>

**Figure 2. Two intuitive accounts of the human circulatory system. They share propositional beliefs and mental model parts, but some propositional beliefs in  $D_a$  are inconsistent with those in  $D_b$ .**

Account  $D_a$  contains beliefs that blood flows from the heart to the rest of the body and back – and nowhere else. Account  $D_b$  contains beliefs that blood *also* flows from the heart to the

<sup>4</sup> We assume – as discussed later in this chapter – that mental models are divisible into reusable components. We simulate these using compositional model fragments, each of which represents a process or conceptual entity (see Chapter 3 for model fragment overview).

lungs and back.<sup>5</sup> Both accounts share some propositional beliefs and mental model pieces, but some propositional beliefs of  $D_a$  are inconsistent with those of  $D_b$ .

An *explanation* is a set of domain knowledge that is joined by logical justifications<sup>6</sup> to explain some phenomenon or event  $m$ , where  $m$  is represented by one or more propositional beliefs in domain knowledge. Domain knowledge, e.g.,  $D_a$ ,  $D_b$ , or any subset thereof, may be in zero or more explanations of phenomena.

Suppose an agent has explained the phenomenon  $m$  = “the body receives oxygen from the blood” with an explanation  $x_a$  that uses the knowledge in  $D_a$ . Suppose also that the agent has decided that  $x_a$  is presently the best account it has of how  $m$  happens. Using terminology from abductive reasoning, we call  $x_a$  the *best explanation* (Peirce, 1958) or *preferred explanation* for  $m$ , since other inferior explanations may exist.

Now suppose the agent reads several sentences of a textbook passage and has acquired the knowledge  $D_b$ , while still entertaining its previous account  $D_a$ . When the agent uses the new knowledge in  $D_b$  to explain  $m$ , a new explanation  $x_b$  is created for  $m$ , and we say that  $x_a$  and  $x_b$  now *compete* to explain  $m$ . Explanations such as  $x_a$  and  $x_b$  are persistent structures, and are used to compartmentalize and contextualize information. This means that the new information  $D_b$  does not replace the existing information  $D_a$ ; rather, the inconsistent beliefs in  $D_a$  and  $D_b$  coexist simultaneously. If the agent compares competing explanations  $x_a$  and  $x_b$  and determines that the new explanation  $x_b$  is better than the presently preferred explanation  $x_a$  (e.g., because it contains new information from a trusted source),  $x_b$  will replace  $x_a$  as the agent’s preferred explanation for  $m$ . This exemplifies part of Claim 3: that the agent constructs explanations and evaluates preferences as a mechanism of change.

---

<sup>5</sup> Neither  $D_a$  nor  $D_b$  is a complete, correct account of the human circulatory system, but both represent mental models of the circulatory system used by middle-school students (Chi et al., 1994a).

<sup>6</sup> We define justifications in Chapter 3.

The decision to replace  $x_a$  to  $x_b$  as the preferred explanation for  $m$  has broad implications for the agent. For instance, if the agent must describe the mechanism of  $m$  on an exam, it can access its preferred explanation  $x_b$  for  $m$  to construct a solution. Alternatively, suppose the agent must explain a novel phenomenon  $m'$  (e.g., the effect of a collapsed lung on the brain's oxygen). To do this, the agent uses similarity-based retrieval (Forbus et al., 1995) to retrieve the relevant phenomenon  $m$ , accesses the best explanation  $x_b$ , and then uses the domain knowledge  $D_b$  within  $x_b$  to explain  $m'$ . If domain knowledge  $D_b$  is used within the preferred explanation  $x_c$  for the new phenomenon  $m'$ , then the set  $D_b$  of domain knowledge now supports the preferred explanations of both  $m$  and  $m'$  and the set  $D_a$  supports neither (though it shares some of the knowledge of  $D_b$ ). Via this system of preferential retrieval and reuse of explanations, beliefs are used and propagated according to whether they participate in preferred explanations. When a belief is no longer a member of a preferred explanation (e.g., the belief “all blood leaving the heart flows directly to the body” in  $D_a$ ), it is effectively inert. This constitutes a mental model transformation. Chapters 6 and 7 describe simulations of mental model transformation via explanation revision.

Claim 3 also states that category revision occurs by the same mechanism of change. Consider a different example: an agent believes that (1) all objects have a quantity  $q$  which has a spatial directional component (e.g., an object can have leftward  $q$ , downward  $q$ , etc.), (2) an object moves if and only if its  $q$  is in the direction of motion, and (3) an object stops moving in a direction if its  $q$  loses that directional component. Consequently,  $q$  is a conflation of weight and momentum, similar to some concepts of force found in the literature (Ioannides & Vosniadou, 2002). Suppose the agent watches a foot strike a large ball and then immediately observes the foot strike a smaller ball, which moves a greater distance. The agent compares the two events,

and determines that the present specification of  $q$  cannot explain the discrepancy in the distances the balls travel. To resolve this explanation failure, the agent considers that  $q$  might be an *acquired* quantity such that one object can transfer some amount of  $q$  to another by touch or collision (rather than shifting the direction of existing  $q$ , previously), and that the transfer rate of  $q$  is inversely proportional to the size of the recipient. This results in a new quantity specification  $q_a$  which is a revision of the previous quantity specification  $q$ .<sup>7</sup>

The agent can use its new quantity specification  $q_a$  to explain why the large and small balls travel different distances. As in the mental model transformation example, the agent formulates new explanations with  $q_a$  rather than  $q$ , and embeds  $q_a$  into preferred explanations of new phenomena. Further, the agent can find *previous* phenomena explained with  $q$  and explain them using  $q_a$ . This process of *retrospective explanation* embeds  $q_a$  in additional preferred explanations and promotes conceptual change. As in the circulatory system example, the previously-existing knowledge loses its likelihood of becoming retrieved and reused, and might eventually become inert.

In our model, category revision and mental model transformation are different types of conceptual change because they involve different types of changes to conceptual knowledge: category revision revises an element (e.g.,  $q$ ) within domain knowledge, and mental model transformation recombines existing elements of domain knowledge (e.g., mental model parts and propositional beliefs) into different aggregates. Importantly, both of these changes are propagated throughout the knowledge base using the same explanation-based process. So while both of these types of conceptual change result in different changes to memory, they share a common propagation mechanisms and underlying memory structure. This completes our discussion of the third claim.

---

<sup>7</sup> Chapter 8 shows how heuristics can be used to revise quantities upon encountering anomalies.



To summarize, constructing and evaluating explanations is the primary mechanism of conceptual change in our cognitive model. We have very abstractly sketched how this occurs, but this does not qualify as a theory or model of conceptual change in itself. In later chapters, we describe the representations and algorithms – including models of explanation construction and explanation evaluation – that underlie this specification. As abstract as it is, our above sketch of the two types of conceptual change does make a number of high-level psychological assumptions that are worth addressing before we discuss the details.

## **1.2 Psychological assumptions of our model of conceptual change**

We summarize our assumptions in Figure 3. Some of these assumptions are supported (*s*) by the literature; these serve as psychological constraints for cognitive modeling. Assumptions that are unsupported (*u*) by the literature serve as psychological predictions of this cognitive model that might be confirmed by later psychological experimentation. Finally, assumptions that are inconsistent (*i*) with the literature are limitations and opportunities for future improvement. We discuss each of these assumptions next.

Type	Assumptions
Knowledge Representation	<ol style="list-style-type: none"> <li>1. Human experts and novices can mentally simulate physical phenomena qualitatively. (<i>s</i>)</li> <li>2. When a person uses a mental model to reason about the world, the object(s) described by the mental model generally correspond to real-world objects. (<i>s</i>)</li> <li>3. People represent causal influences between quantities in their intuitive knowledge about the world. (<i>s</i>)</li> <li>4. Regardless of how they are organized within theories and explanations, human mental models can be represented as reusable parts. (<i>s</i>)</li> <li>5. People store mental models in long-term memory. (<i>s</i>)</li> <li>6. People can learn and reason with propositional beliefs. (<i>s</i>)</li> </ol>
Memory & Access	<ol style="list-style-type: none"> <li>7. People can evaluate competing explanations for a single phenomenon. (<i>s</i>)</li> <li>8. People can believe two inconsistent beliefs simultaneously when those beliefs are used to explain different phenomena. (<i>s</i>)</li> <li>9. After explaining a phenomenon, people generally retain the best explanation for the phenomenon in long-term memory, but may not discard other explanations. (<i>u</i>)</li> <li>10. When explaining a novel phenomenon, people often retrieve a similar, previously-understood phenomenon to aid in explanation. (<i>s</i>)</li> </ol>
Learning	<ol style="list-style-type: none"> <li>11. People use analogy to generalize the common structure of observations. (<i>s</i>)</li> <li>12. People can revise the ontological properties of a quantity concept. (<i>s</i>)</li> <li>13. People do not immediately replace concepts through conceptual change; throughout the process, people have access to both the old and new knowledge (e.g., quantities and mental models). (<i>s</i>)</li> <li>14. People can change how they explain a phenomenon. (<i>s</i>)</li> <li>15. The cognitive processing required to transition away from a misconception is qualitatively proportional to how pervasively the misconception was previously used in explanations. (<i>u</i>)</li> <li>16. By (15), some misconceptions are more resilient to change than others. (<i>s</i>)</li> </ol>

**Figure 3. High-level psychological assumptions of our cognitive model, organized by type. Each is labeled where supported by (*s*) or unsupported by (*u*) the literature.**

### 1.2.1 Assumptions about knowledge representation

Toulmin (1972) argued that the term “concept” is pervasively used and ill-defined in the literature, and this complaint is still warranted today. As Figure 1, illustrates, we have a very specific representation of conceptual knowledge, using existing knowledge representation formalisms in AI. Further, our claim that human mental models can be simulated with compositional model fragments has been argued previously (e.g., Forbus & Gentner, 1997). Still, we review each of the related assumptions, since the mental model literature has not reached consensus on knowledge representation.

In support of assumption #1: *use of qualitative reasoning*, there is evidence that novices and experts alike often reason with incomplete and imprecise qualitative knowledge, especially in situations of informational uncertainty (Trickett & Trafton, 2007). This supports our choice of using compositional qualitative models to simulate human mental models. We describe qualitative reasoning in more detail in Chapter 3.

The term “mental models” (Gentner & Stevens, 1983; Gentner, 2002) has been widely used to describe representations of domains or situations that support everyday explanation and prediction. Nersessian (2007) provides generally-accepted criteria for psychological mental model-based reasoning:

- It involves the construction or retrieval of a mental model.
- Inferences are derived through manipulation of the mental model.

Vosniadou & Brewer (1994) note additional characteristics of mental models:

- The observable and unobservable objects and states of the world that a mental model represents are often analogs of real-world objects. (Supports assumption #2: *entities and states in mental models have real-world correspondences.*)
- Mental models provide explanations of physical phenomena.
- Many mental models may be manipulated mentally or “run in the mind’s eye” to make predictions about the outcomes of causal states in the world.

One representation distinction noted by Markman and Gentner (2001) is between *logical mental models* and *causal mental models*. In the logical mental model account (e.g., Johnson-Laird, 1983), mental models are logical constructs in working memory. In this view, mental models are constructed on-the-spot, involving only knowledge in working memory about the local problem-at-hand. This approach has been criticized for failing to simulate human reasoning that is captured by propositional reasoning (Rips, 1986). This definition of mental models is inconsistent with assumption #5: *mental models in LTM*.

In the causal mental model account (e.g., Gentner & Stevens, 1983), the entities and quantities of a mental model correspond to observable and unobservable entities and quantities in a causal system (supporting assumption #2: *entities and states in mental models have real-world correspondences*). Further, causal mental models draw on long-term domain knowledge (supporting assumption #5: *mental models in LTM*). In this dissertation, we use the term “mental model” to refer to this causal account of mental model.

In support of assumption #3: *representing quantity influences*, there is evidence that even infants have knowledge about the relationship between quantities. For example, 6.5-month-old infants look reliably longer – indicating a violation of expectation – when a small object *A* strikes

a second object *B* and causes *B* to roll farther than when a large object *C* hits the same object *B*. This suggests that infants understand an *indirect influence* between quantities: the distance something travels is qualitatively proportional to the size of the object that strikes it (Baillargeon, 1998). It is safe to assume that humans have a tendency to represent influences between quantities, even prior to formal instruction, but not prior to experience. We describe direct and indirect influences in depth in Chapter 3.

Our assumption #4: *piecewise mental model representation* is a key argument of compositional accounts of mental models (Collins & Gentner, 1987) and of the *knowledge in pieces* (hereafter KiP) view of conceptual change (diSessa, 1993; diSessa et al., 2004). KiP claims that conceptual reasoning involves the coordination of various *phenomenological primitives* which include rules, constraints, and qualitative proportionalities such as *larger objects have greater momentum*. Under KiP, conceptual change involves revising a piece of knowledge or recombining them to generate new explanations.

The plausibility of assumption #4 is not limited to the KiP perspective. For example, researchers who oppose KiP and advocate a more coherent account of human mental models (e.g., Vosniadou & Brewer, 1992; Vosniadou, 1994; Ioannides & Vosniadou, 2002) describe the existence of *synthetic mental models*. In this coherence-based account, synthetic mental models are the result of partially revising an intuitive (i.e., pre-instructional) mental model to accord with scientific knowledge. One example of a synthetic model found by Vosniadou & Brewer (1992) is a flat, disc-shaped earth, formed by students who assimilate the knowledge “the earth is round” into an intuitive model of a flat, rectangular earth. If we can identify components of this synthetic model as intuitive (e.g., the flatness of the disc-earth) and other aspects as instructional

(e.g., the roundness of the disc-earth), then we can say that even though human mental models might be stored coherently, they are at least plausibly represented as smaller components.

Representation assumption #6: *people reason with propositional beliefs* is widely (though not universally) accepted in cognitive science (Forbus & Gentner, 1997; Chi, 2008; Vosniadou, 1994; but see Glenberg et al., 1999; Thelen and Smith, 1994). This is supported by studies of deductive reasoning (e.g., Rips, 2001) and accounts of conceptual change (e.g., Chi, 2008; Vosniadou, 1994; diSessa, 1993). This does not mean that propositional beliefs are always easy to change; to the contrary, Vosniadou (1994) argues that *presuppositions* – prevalent propositional beliefs such as “things that are unsupported from beneath fall down” – are the most difficult to change.

### 1.2.2 Assumptions about memory and knowledge organization

We now discuss psychological assumptions about how knowledge is evaluated and organized in long-term memory.

Assumption #7: *evaluating competing explanations* is supported by the literature. Chinn et al. (1998) propose that everyday people evaluate explanations based on the criteria of empirical accuracy, scope, consistency, simplicity, and plausibility, and scientists evaluate scientific explanations by the additional criteria of precision, formalisms, and fruitfulness. Lombrozo (2011) mentions additional *explanatory virtues* by which people judge explanations, including coverage of observations, goal appeal, and narrative structure.

There is evidence in the literature for assumption #8: *simultaneous inconsistent beliefs*. For example, Collins and Gentner (1987) found that novices often use mutually inconsistent mental models of evaporation and condensation to explain different phenomena. While novice

explanations are locally consistent for explaining individual phenomena (e.g., hot water evaporating in a refrigerator, seeing your breath in the winter) they may be globally inconsistent. Since these inconsistent mental models are narrowly compartmentalized by phenomena, the learner may never realize these inconsistencies (Gentner, 2002).

Our model assumes that people store explanations for phenomena – including justification structure – in long-term memory (assumption 9). This is probably a case where the model's assumptions are too strong. Other theories of conceptual change suggest that explanations are organizational structures (e.g., Carey, 1985), but it seems unlikely that people retain all of the justification structure of their explanations. Evidence suggests that if people do retain justifications for their beliefs (and by extension, the entire explanation(s), according to assumption #9) they tend to retain a belief even after the supporting evidence is discredited. Ross and Anderson (1982) discuss several experiments that (1) convinced people of a belief (e.g., the professional performance of firefighters positively or negatively correlates with their score on a paper and pencil test) and then (2) debriefed the subject to communicate that the initial evidence was fictitious – and that in fact, the opposite was true. In these studies, the subject retained significant confidence in the belief after the evidence was discredited, suggesting that evidence is not required for retaining a belief. In a similar study (Davies, 1997) people either read high-quality explanations for the outcomes of an event or constructed explanations for themselves for the same outcomes, based on the same evidence. After all of the evidence was discredited, subjects who constructed explanations for themselves were significantly more likely to retain the unsupported belief than those that read a high-quality explanation.

Since people do not always rely on the evidence for their beliefs during everyday belief revision, they might not encode all of the justifications between beliefs and supporting evidence. Harman (1986, pp. 41) argues that “[i]t stretches credulity to suppose people always keep track of the sources of their beliefs but often fail to notice when the sources are undermined.” This is a philosophical appeal to the simplest explanation, which is part of a larger debate in the philosophy and belief revision literature between the *foundations* theory and the *coherence* theory. Though philosophical appeals to this question do not provide us with empirical evidence for our assumption, they help illustrate the dilemma.

According to the foundations theory (e.g., Doyle, 1992), justifications for beliefs are retained, and a rational agent holds a belief if and only if it is justified.<sup>8</sup> If all justifications for a belief are invalidated, that belief is invalidated, and the justifications *it* supports are also invalidated, resulting in a possible chain-reaction. Conversely, under the coherence theory (e.g., Gärdenfors, 1990), justifications for beliefs are not retained in memory – if the agent is no longer justified in believing something (i.e., there is no more evidence), the belief is still retained insofar as it is consistent with other beliefs. Put simply, the foundations theory states that beliefs are *held* only if there is rationale, and the coherence theory states that once a belief is held, it is only *removed* if there is rationale.

Our cognitive model does not strictly adhere to the foundations theory, since beliefs are not necessarily retracted when they lose support (i.e., they may become assumptions if they are used to support other beliefs), but it does rely on justification structure of explanations to organize beliefs. Other systems that record justification structure (e.g., Doyle and Wellman, 1990) also retain unjustified beliefs when convenient.

---

<sup>8</sup> Belief revision according to the foundations theory is exemplified by Truth Maintenance Systems (Forbus & de Kleer, 1993), discussed in Chapter 3. AI approaches that track justifications of knowledge generally encode justifications for all premise beliefs. Consequently, observations are intrinsically justified.



Since we have no hard evidence to support assumption #9, our model might rely too heavily on the presence of explanations in long-term memory. We describe some ideas for altering the model to remove this assumption in section 9.4

There is indirect evidence in the literature for assumption #10: *retrieval of a similar, understood phenomenon*. During problem solving, people are often reminded of prior problems; however, these reminders are often based on surface-level similarities between problems rather than deeper relational similarities (Gentner, Ratterman, & Forbus, 1993; Ross, 1987). On the rare occasions that they retrieve a useful analog in a distant domain, people can use these cases via analogy to the present problem to find a solution (Gick & Holyoak, 1980). There is evidence that people have some success in retrieving and utilizing similar problems in the domains of mathematics (Novick, 1988) and computer programming (Faries & Reiser, 1988). It is therefore a safe assumption that people are reminded of similar phenomena when faced with a new phenomenon to explain, especially when they have surface-level similarity. This still allows for the possibility that nothing may be retrieved, e.g., when episodic memory is empty or when no previously-encountered phenomena are similar. The simulation described in Chapter 8 uses heuristics to generate new domain knowledge in these instances.

### **1.2.3 Assumptions about learning**

Our claim that people can induce mental models from observations assumes that people use analogy to generalize (assumption 11). It also makes assumptions regarding how people represent their observations, which we address later. There is substantial evidence that both adults and children use analogical generalization to learn categories and relationships over very few examples. For instance, 4-year-olds can learn the abstract relational categories *monotonicity*

and *symmetry* with only a few examples, if done correctly, which is elegantly explained by analogical generalization (Kotovsky & Gentner, 1996). Further, Gentner and Namy (1999) found that when 4-year-olds are provided a single example of a nonsense category such as a “dax,” and asked to find another dax, they choose a perceptual (i.e., surface-level) match; however, when given two training examples and encouraged to compare, they pick a conceptual (i.e., relational) match. This suggests that the act of comparing as few as two examples can induce a new category hypothesis, which is consistent with analogical generalization.

Our model assumes that people can make ontological revisions to their concepts (assumption 12). Ontological revision is a central component of Chi’s (2005; 2008) theory of conceptual change. Chi calls ontological revision a *categorical shift*, whereby a category such as “Whale” changes lateral position in a hierarchical ontology of categories, e.g., from a subordinate position of “Fish” to a subordinate position of “Mammal.” The more distant the initial and final position of a concept, the more difficult the conceptual change. Two notable examples are as follows: (1) shifting “Force” from its intuitive position under “Substance” (Reiner et al., 2000) to a lateral “Constraint-based interaction” position (Chi et al., 1994b); and (2) shifting “Diffusion” from beneath “Direct process” to beneath “Emergent process” (Chi, 2005). Our model does not rely on these specific ontologies, but it does assume that people are capable of making ontological changes, and this assumption seems safe.

Since our model of conceptual change involves incrementally transitioning between theories, we rely on assumption #13: *theories are not immediately replaced*. For example, it cannot be the case that learning a new and credible theory of dynamics causes a person to immediately forget the inconsistent beliefs and models of a previous theory of dynamics. AI algorithms for coherence-based belief revision (e.g., Alchurron et al., 1985) immediately remove

inconsistent beliefs in this fashion. Similarly, dependency-directed backtracking algorithms for truth maintenance (e.g., Doyle, 1979; Forbus & de Kleer, 1993) immediately retract assumptions to retain consistency. Since we assume that people can hold contradictory beliefs (assumption 8), these algorithms are not used in our conceptual change model.

The literature supports the assumptions that competing theories can coexist, psychologically. In their constructivist view of conceptual change, Smith, diSessa, and Roschelle (1994) note that as people accrue theories, they evaluate them with respect to their effectiveness in understanding and manipulating the world. Under this view, nonscientific theories can be used productively even when scientifically-correct theories are available. Similarly, students often learn to use quantitative Newtonian theories of force while still operating with their qualitative misconceptions of force (Clement, 1985; Hestenes et al., 1992). The Newtonian laws, e.g.,  $F = ma$  can also be used for qualitative reasoning. For instance, all else being equal, increasing mass must increase force (i.e., force is qualitatively proportional<sup>9</sup> to mass) and increasing force must increase acceleration (i.e., acceleration is qualitatively proportional to force). The predictions of this qualitative Newtonian theory of force are inconsistent with most students' intuitive qualitative models of force. Despite their joint applicability, students might contextualize Newtonian and intuitive models of force separately, so that Newtonian models are used in quantitative classroom problem-solving and intuitive models are used in everyday qualitative reasoning contexts. This micro-contextualization of mental models is not a new idea; Collins and Gentner (1987) suggest that this is the reason novices are able to reason with inconsistent knowledge, often without detecting an inconsistency.

As described above, our model of conceptual change involves incrementally shifting phenomena from explanations that use a superseded theory to explanations that use a preferred

---

<sup>9</sup> Qualitative proportionalities are described in greater detail in section 3.2.

theory. Consequently, we make the assumption #14: *phenomena can be re-explained*. This is not a contentious claim – studies that contain a pretest and posttest to measure learning (e.g., about the human circulatory system in Chi et al., 1994a) or an interview during which explanations change (e.g., about the changing of the seasons in Sherin et al., in press) demonstrate clearly that people can change their explanations for phenomena. This may not be sufficient to show that people retain all of the justifications for their explanation (assumption #9), but they do associate the phenomenon with new – or at least, different – supporting knowledge.

Since our computational model relies on the gradual shift of explanatory support, it follows that the more explanations include a theory, the more computations are necessary for the agent to transition away from said theory. In other words, we predict that the more pervasive a misconception is, the more processing is required to overcome it (assumption #15). There is no direct support of this in the literature, but this is consistent with the idea that productive theories are more pervasive and robust to change (Smith, diSessa, and Roschelle, 1994).

If we assume that some misconceptions require more processing to overcome than others (assumption #15) then we arrive at assumption #16: *some theories are more resilient to change*. As mentioned above, Vosniadou (1994; Vosniadou & Brewer, 1992, 1994) makes a distinction between mental models and presupposition beliefs that constrain these mental models. For example, a mental model of a flat earth is constrained by the presupposition “things that are unsupported from beneath fall down.” In Vosniadou’s theory, these presuppositions are more resilient to change than the mental models they constrain. Further, de Leeuw (1993) and Chi (2000) argue that the perseverance with which a belief is held increases with the number of

consequences the belief has in a network therein.<sup>10</sup> In our model, these networked consequences of a belief correspond roughly to the explanations that include said belief. Our definition of *theory* includes a set of beliefs, so this supports the assumption that theories vary in their resilience to change.

Researchers have also characterized how people resist changing their beliefs. People use evasive strategies called *knowledge shields* (Feltovich et al., 2001) to ignore anomalous data, and they use other strategies such as rejecting, reinterpreting, excluding, and holding knowledge in abeyance (Chinn & Brewer, 1993; 1998) to resist change. In the event that people do revise their beliefs, they frequently make minimal changes to their present theory rather than adopting a new theory in its entirety (Posner et al., 1982; Chinn & Brewer, 1998). All of the simulations described below are biased toward minimizing changes. For example, Chapter 7 describes simulation trials that learn humanlike misconceptions by choosing to use concepts (e.g., “heart”) known prior to instruction over other concepts (e.g., “left-heart”) that were acquired by formal instruction. Since the focus of this dissertation is conceptual change, we are more interested in simulating the successful – albeit minimal – revision of beliefs rather the avoidance of belief change; however, modeling avoidance strategies is an interesting opportunity for future work.

To support the claims of this dissertation, we have developed a model of conceptual change, implemented the model on a cognitive architecture, and conducted four simulation experiments to compare the trajectory of models that the system undergoes to the trajectory of mental models of human learners. Our computational model is described in Chapter 4, and is a novel contribution of this dissertation. The only aspects of our model that are not novel contributions are described in Chapter 3, our discussion of background AI technologies.

---

<sup>10</sup> It is unclear whether “consequences” refer to logical entailments (in the philosophical coherence-based view of belief revision) or justifications (in the philosophical foundations view of belief revision) supported by a belief. Regardless, this supports the assumption that some beliefs are more resilient to change than others.

In the next chapter we describe other theories of conceptual change from the cognitive science literature and discuss areas of contention between them. A comparison of our model with these previous models is best done after our simulation results are presented, and hence is postponed until Chapter 9.

## Chapter 2: Other theories of conceptual change

One aim of the cognitive model presented in this dissertation is to provide insight into the cognitive processes underlying human conceptual change. This warrants a discussion of existing theories of conceptual change and the areas of dispute that our model might help explicate.

None of the conceptual change theories we discuss have computational models that capture the full spectrum of belief changes they describe.<sup>11</sup> Consequently, some speculation is necessary for determining each theory's constraints on knowledge representation, memory organization, and revision mechanisms.

Despite the consensus that *concepts* are the granularity of change in conceptual change, different theories of conceptual change make different assumptions regarding what a concept is and how they change (diSessa and Sherin, 1998). No theory uses the word “concept” exactly as any other theory does or exactly as we do in our cognitive model – in fact, we try to avoid this vague term. Unfortunately, we must use “concept” when discussing other theories to avoid making over-specific assumptions about knowledge representation, since the theorists' definitions of “concept” may be intentionally abstract or noncommittal.

Ideally, we could compare our model of conceptual change with other computational models that implement these four theories: they could learn from the same training data and we could monitor their progress over time using the same testing data. Unfortunately, since none of these theories have computational models that capture the full spectrum of belief changes they describe, this is not feasible. The other possibility is to modify our model to reflect the different

---

<sup>11</sup> INTHELEX (Esposito et al., 2000a; Vosniadou et al., 1998) has been used to model aspects of conceptual change in learning the meaning of “force” using logical theory refinement; however, the system is given multiple representations of “force” concept (e.g., “internal” force and “acquired” force) from the start, and does not invent and transition between representations spontaneously as children do, according to Ioannides and Vosniadou (2002).

aspects of these theories. This is not feasible, since the underlying algorithms and knowledge representations have not been specified for these theories. Ultimately, we must compare our model to these four theories by abstracting the assumptions and behaviors of our model into a psychological theory of conceptual change, and then comparing the theories at that level. We save this discussion for Chapter 9, after we have presented the data from our simulations.

This chapter begins by describing four theories of human conceptual change that aim to explain how people adopt new beliefs in the presence of conflicting beliefs. For each theory, we discuss its underlying assumptions about knowledge representation, memory organization, and mechanisms of change. After discussing these theories of conceptual change, we discuss some notable areas of divergence and disagreement.

## **2.1 Four theories of conceptual change**

The conceptual change theories we discuss include the theory-theory of conceptual development, framework theory, categorical shift, and knowledge in pieces. Each theory makes different commitments to the representation of categories and mental models, the organization of this knowledge in the mind, and the mechanisms that carry out change.

### **2.1.1 Carey's theory-theory of conceptual development**

We begin by discussing Susan Carey's (1985; 1988; 2009) theory of conceptual change. Carey's theory is characterized by a strong appeal to the history of science to draw similarities between conceptual change in children and in the scientific community. It also relies on Kuhn's (1962) notion of *incommensurability* between conceptual systems. Incommensurability is a relation that holds between the languages of two theories. Two conceptual systems (i.e., theories with



propositional beliefs, categories, and models)  $CS_1$  and  $CS_2$  are incommensurable if  $CS_1$  contains concepts that are incoherent from the point of view of  $CS_2$ . That is, the beliefs, laws, and explanations that can be stated in  $CS_1$ 's language cannot be expressed in the language of  $CS_2$ . The presence of concepts in  $CS_1$  that are merely absent in  $CS_2$  is not sufficient for incommensurability.

For an example of incommensurability, consider Jean Buridean's theory of projectile dynamics (based heavily on Aristotelian dynamics) with respect to Newtonian projectile dynamics. Buridean and Newtonian dynamics use different vocabularies – Buridean uses the concept of *impetus*, and Newton uses the concept of *force*. The Buridean concept of impetus is proportional to velocity, so an impetus in the direction of motion *sustains* an object's velocity. Newtonian net force is proportional to acceleration, so a non-zero net force in the direction of motion *increases* an object's velocity. Also, an object moving at constant velocity has a constant impetus (i.e., the impetus is not weakened by gravity or air resistance) in Buridean theory, but it has a zero net force in Newtonian theory. A final point of contrast is the motion of bodies on circular paths. Buridean's theory states that circular impetuses sustain the circular motion of celestial bodies. In some ways, this is a simpler explanation than accounting for the tangential velocity of orbiting bodies with inward acceleration due to the curvature of space-time. Carey's examples of incommensurability include other historical examples (e.g., the source-recipient theory of heat versus the caloric theory of heat) and developmental examples (e.g., theories of physics with and without weight differentiated from density).

Under Carey's theory, conceptual change involves a shift from a conceptual system  $CS_1$  to an incommensurable conceptual system  $CS_2$ . Both conceptual systems are internally coherent, stable, and symbolically represented. The difficulty of achieving conceptual change in some

domains, e.g., learning to differentiate weight from density, stems from this incommensurability. When novices and experts hear “weight,” they understand something different, and the corresponding novice and expert ideas are mutually incoherent. This is an obstacle for effective communication and formal instruction. Since children must acquire the scientific account CS<sub>2</sub> via social processes, incommensurability makes conceptual change difficult.

The process of conceptual change must therefore create representations for CS<sub>2</sub> that are qualitatively different from those in CS<sub>1</sub>. Carey (2009) argues that children perform *Quinian bootstrapping* to achieve this. Quine (1960) describes bootstrapping using a metaphor: you use a ladder to build a platform in a conceptual system until the platform is self-sustaining, and then you kick the ladder out from under. In the case of historical and psychological conceptual change, the symbols that represent concepts (e.g., *weight* and *density*) are used as placeholders for developing a new conceptual system CS<sub>2</sub>. Processes such as analogy (e.g., Gentner et al., 1997), model-based thought experimentation (e.g., Nersessian, 2007), and abduction are used to integrate new knowledge and support observations using these placeholder symbols. In this manner, placeholder concepts are learned together and gain meaning relative to each other. This bootstrapping process is iterative, and through successive rounds of analogy, abduction, and model-based reasoning, the concepts in CS<sub>2</sub> acquire meaning and are used to explain real-world phenomena.

### **2.1.2 Vosniadou’s framework theory**

Like Carey’s theory-theory of conceptual development, Vosniadou’s (2002; 1994; Vosniadou and Brewer, 1992; 1994; Ioannides and Vosniadou, 2002) theory posits that novices have an internally coherent intuitive understanding of the world that is subject to modification and radical

revision. In place of Carey's conceptual systems, Vosniadou uses the term *framework theories*. Children's framework theories are coherent explanatory systems, but they lack characteristics of scientific theories such as systematicity, abstractness, social nature, and metaconceptual access (Vosniadou, 2007; Ioannides and Vosniadou, 2002). Embedded within framework theory are *specific theories* about phenomena (e.g., the day/night cycle) and entities (e.g., the earth). Specific theories are also referred to as *specific explanations* (Ioannides and Vosniadou, 2002). Finally, embedded within the framework theory and specific theories are mental models. The embedded nature of knowledge refers to the direction of constraint: the framework theory constrains the specific theories/explanations, which in turn constrain the mental models (Vosniadou, 2002).

Framework theories contain *presuppositions*, which are propositional beliefs that are learned from observations and cultural influences. Each presupposition places consistency constraints on the specific theories embedded within the framework theory. In this fashion, presuppositions limit the space of allowable specific theories, and indirectly, the space of allowable mental models. For example, the presupposition "unsupported objects fall down" affects the specific theory and mental model of the earth, since a spherical earth with people standing on the "bottom" would contradict the presupposition. It is assumed that changing a specific theory (e.g., of the shape of the earth) is easier than retracting presuppositions, provided the new specific theory is consistent with existing presuppositions.

In Vosniadou's theory, the main difficulty of conceptual change is that students frequently assimilate aspects of a scientific explanation into their flawed framework theory without sufficiently revising their presuppositions. In these cases, learners either (1) do not notice the contradictions between the new information and their presuppositions and explanations, or (2)

they notice contradictions and only make partial (i.e., insufficient) changes to their presuppositions and explanations. Partial revision of a framework theory can produce new misconceptions and *synthetic models* (Vosniadou and Brewer, 1992; 1994; Ioannides and Vosniadou, 2002), which are incorrect mental models that incorporate both intuitive and scientific components. Consider integrating the belief “the earth is round” into a framework theory that contains the “unsupported objects fall down” presupposition with a mental model of the earth as a flat rectangle. Since presupposition theories are more resilient, the mental model of the earth is the easiest component to revise, and the earth may be conceived of as a flat cylinder, a flattened sphere, or even a hollow sphere with a flat surface inside (Vosniadou and Brewer, 1992). The mental model of the earth is thereby constrained by the presupposition, and the learner must revise this presupposition to acquire the correct mental model of the earth.

Changing a framework theory is a gradual process, driven by observation, explanation, and formal education. Throughout the process of learning science, aspects of scientific theories are assimilated into the theories/explanations embedded within the student’s framework theory, as well as into the framework theory itself. This yields a series of synthetic models which approach the correct scientific theory.

### **2.1.3 Chi’s categorical shift**

Chi and colleagues (Chi, 2008; 2005; 2000; Reiner et al., 2000; Chi et al., 1994b) distinguish between three different types of conceptual change: (1) categorical shift; (2) mental model transformation; and (3) belief revision. All three types of conceptual change require that some existing knowledge is retracted or revised; otherwise, this would constitute gap-filling, enrichment, or *tabula rasa* knowledge acquisition. We discuss each of these types of change

according to Chi's theory, including the type of knowledge affected and the mechanism of change.

*Categorical shift* was briefly discussed in the previous chapter. It involves changing a category's lateral position in a hierarchy of categories. Chi's theory assumes the existence of multiple, disconnected *ontological trees* with multiple levels of inheritance. For instance, Chi (2008) identifies three ontological trees: (1) "Entities" which has subordinate branches "Concrete Objects" and "Substances;" (2) "Processes" which has branches "Direct," and "Emergent," and (3) "Mental States" with branches "Emotion" and "Intention." Each tree and level in the hierarchy ascribes *ontological attributes* to subordinate categories, e.g., a lamp (under the "Artifacts" branch of the "Entities" tree) can be broken and a hug (under the "Events" branch of the "Processes" tree) can be a minute long. All else being equal, the greater the lateral distance between two categories, the more their ontological attributes differ. This distance is an important consideration for Chi's theory, because shifting a category from one place in the hierarchy to another involves changing ontological attributes – and the greater the distance, the greater the change. For example, "Fish" and "Mammals" categories both share the close ancestor category of "Animals" under the "Entities" tree. These categories are much closer than "Substances" (under the "Entities" tree) is to "Constraint-Based Interactions" (under the "Processes" tree). Shifting "Whale" from "Fish" to "Mammals" is easier (i.e., less ontological attributes must change) than shifting a category such as "Force" from "Substances" (Reiner et al., 2000) to "Constraint-Based Interactions." Categorical shifts are incommensurate, according to Carey's (1985) definition of incommensurability (Chi, 2008).

In Chi's theory, *belief revision* occurs at the granularity of propositional beliefs, when new information is logically inconsistent with prior beliefs. For example, the belief "the heart

oxygenates blood” is inconsistent with the new information “only the lungs oxygenate blood.” When this occurs, students can retract the existing belief, adopt the new information, and continue looking for inconsistencies. In reality, students generally encounter information that conflicts less directly with their existing beliefs, such as “the lungs oxygenate blood” (i.e., still logically permitting the heart to oxygenate blood also), but they still achieve successful belief revision even through indirect, implicit conflict (Chi, 2008).

The third type of conceptual change in Chi’s theory is *mental model transformation*, which is a special case of belief revision. In Chi’s framework, mental models are organized groups of propositional beliefs which can predict changes and outcomes in a situation or system such as the human circulatory system. When a mental model is *flawed*, it is internally coherent but generates incorrect explanations and predictions. Two mental models (e.g., a flawed and a correct model) are in conflict when they make mutually inconsistent predictions and explanations, even though the beliefs that comprise the mental models might not be explicitly contradictory. Mental models are ultimately transformed by the revision of the beliefs that comprise the mental model. For this to occur, new information must be in explicit or implicit conflict with the beliefs of the mental model, according to the above description of belief revision. Some false beliefs are more “critical” than others (Chi, 2008) in that they discriminate between a flawed and correct model. For example, the false belief “the heart oxygenates the blood” is more critical to explaining and predicting the behavior of the circulatory system than the false belief “all blood vessels have valves.”

These accounts of belief revision and mental model transformation do not involve incommensurability, as defined by Carey (2009). This is because a mental model shares the same symbolic vocabulary before and after its transformation, even though entities may be added

or removed. This assumes that no categorical shift occurs during mental model transformation. Only categorical shifts involve incommensurability, since the vocabulary changes (i.e., categories gain and lose ontological attributes).

#### **2.1.4 diSessa's knowledge in pieces**

The Knowledge in Pieces (KiP; diSessa, 1988; 1993) view argues that intuitive knowledge consists of a multitude of inarticulate explanatory phenomenological primitives (*p-prims*) which are activated in specific contexts. P-prims are phenomenological in that (1) they provide a sense of understanding when they are evoked to explain or interpret a phenomenon and (2) they provide a sense of surprise when they cannot be evoked to explain a situation or when their predictions are inconsistent with reality. They are primitive in that they are generally invoked as a whole and they need no justification.

P-prims are not systematic enough to be described individually or collectively as a coherent theory (diSessa et al., 2004). Furthermore, a student may operate with an incoherent set of p-prims – that is, his or her p-prims may make conflicting predictions about a situation, similar to Chi's (2008) account of conflicting mental models. This is in direct disagreement with the coherent nature of Carey's conceptual systems and Vosniadou's framework theories.

A person or an AI system with incoherent conceptual knowledge may seem unlikely or unproductive to some, but according to KiP, each piece of knowledge is highly contextualized with respect to its applicability in the real world (diSessa et al., 2004). This allows people to provide coherent explanations for individual phenomena despite global inconsistency.<sup>12</sup> If a

---

<sup>12</sup> Collins and Gentner (1987) provide empirical evidence that novices can narrowly contextualize inconsistent mental models to achieve internally consistent explanations, but their account of mental models (see Gentner and Stevens, 1983) is not committed to fragmentation or p-prims, according to the knowledge in pieces perspective.

novice generates a coherent explanation, it is an effect of knowledge contextualization and of the process of explanation construction; it is not a hard constraint on how knowledge is organized in memory.

Since KiP does not involve coherent theories or conceptual systems, what constitutes misconceptions and conceptual change? Smith, diSessa, and Roschelle (1993) argue that the standard model of misconceptions – that students hold flawed ideas which are replaced during instruction – conflicts with the premise of constructivism that students build more advanced knowledge from existing understandings. KiP emphasizes the continuity from novice to expert knowledge the presence of intuitive knowledge within expert understanding (Sherin, 2006). Consequently, KiP focuses on knowledge refinement and reorganization rather than replacement. Minstrell's (1982, 1989) KiP account of conceptual change involves the recombination of explanatory primitives and reuse in different contexts. Similarly, diSessa (1993) describes how the contexts and priorities of p-prims can be altered to change how learners construct explanations and predictions in future situations.

Under KiP, the difficulty of conceptual change is a factor of how productive a piece of knowledge is within a given context. Suppose a learner has previously predicted and understood the world using the kinematic “blocking” p-prim (diSessa, 1993) whereby an object such as a brick blocks a moving object without any sense of effort or strain (e.g., the brick does not visibly move, bend, or compress). The more productively “blocking” has been at explaining and predicting within a class of phenomenon (e.g., putting objects atop rigid surfaces, thus preventing the object from moving further downward), the more difficult it will be to assign other knowledge besides “blocking” (e.g., of normal forces) to be evoked in this context.



## **2.2 Divergence and disagreement**

All of the above theories aim to explain documented examples of conceptual change, so there is considerable consensus about the principles and constraints of conceptual change. There are also many points of contention among the four theories outlined above. We discuss four topics that lack consensus which are especially relevant to our cognitive model: (1) what counts as conceptual change; (2) revision versus addition (3) the effect of explaining; and (4) the source of coherence. We discuss these topics with regard to our model in Chapter 9, after we have described the simulations that exemplify our model's behavior.

### **2.2.1 What counts as conceptual change**

Carey (2009) argues that incommensurability is a necessary condition for conceptual change. This necessarily involves creating new primitives, symbols, and relationships that were not coherently describable in the language of the existing conceptual system. Requiring incommensurability sets Carey's theory apart from the other theories.

Chi's (2008) account of conceptual change includes categorical shift (i.e., change of the incommensurable sort) and also commensurable changes such as mental model transformation and belief revision. Similarly, Vosniadou (1994; Vosniadou and Brewer, 1992; 1994; Ioannides and Vosniadou, 2002) considers the revision of mental models a type of conceptual change. Changing the presuppositions of a framework theory – a type of belief revision – is a key operation in Vosniadou's theory of conceptual change.

Also in disagreement with Carey, diSessa (2006) argues against the necessity of incommensurability within conceptual change. Collecting and coordinating elements of

knowledge is the mechanism of conceptual change for KiP, so incommensurability is not a worthwhile distinction.

This particular point of contention concerns terminology rather than human cognitive processes. Carey (2009) states clearly that, “[c]onceptual change’ means change in individual concepts” (pp. 354), but the other theories – most notably, Chi’s – include other manners of non-monotonic belief revision (i.e., removing beliefs to accommodate new information). We include mental model transformation in our definition of conceptual change, as described in Chapter 1. We also include category revision, which abides by Carey’s definition of conceptual change.

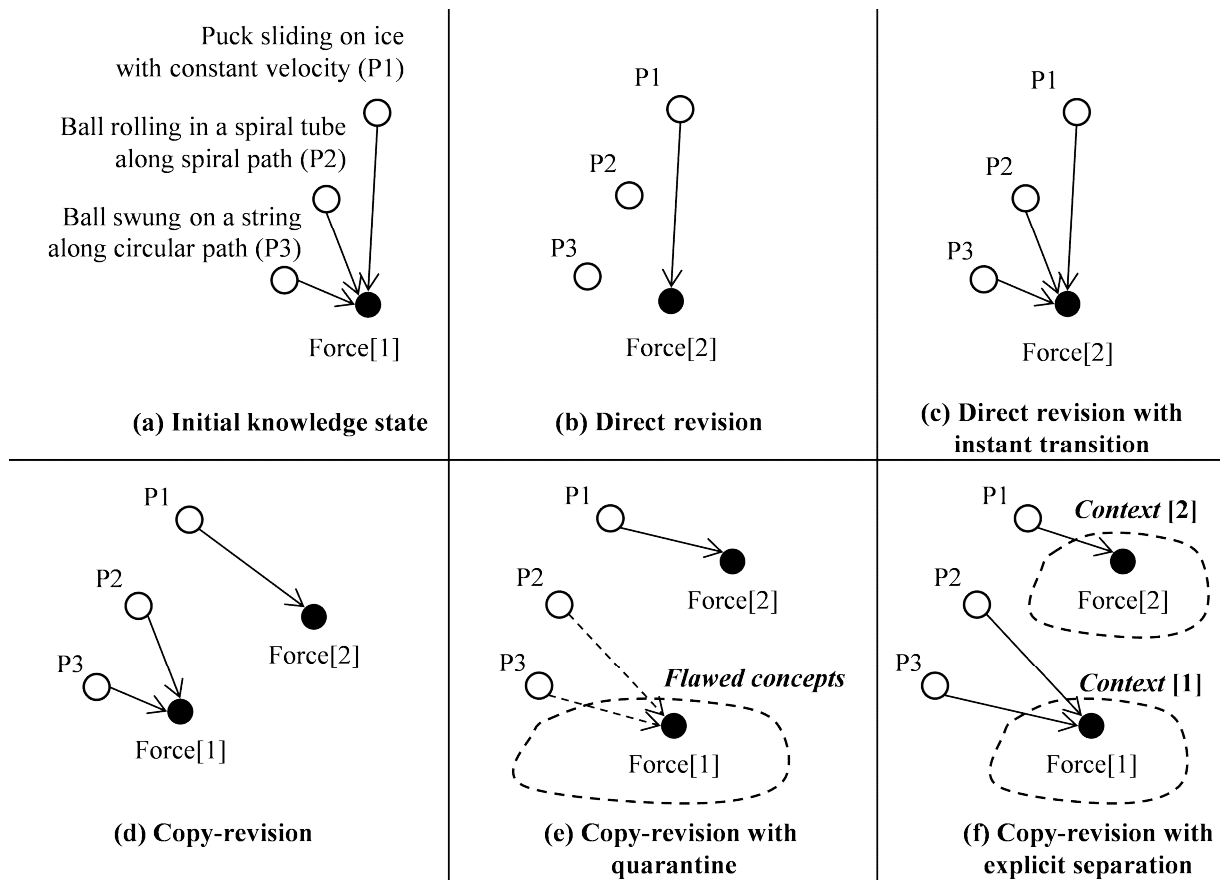
### **2.2.2 Revision versus addition**

There is a deep but subtle distinction between these theories of conceptual change that has not, I believe, been given sufficient attention. It concerns the revision of information in memory. Consider the following example of conceptual change: a student is learning Newtonian dynamics. She generally operates with a flawed account of force, in that it is substance-like (Reiner et al., 2000), impetus-like (Ioannides and Vosniadou, 2002), or it includes the “force-as-mover” p-prim (diSessa, 1993). Consequently, she generally believes that motion implies the existence of a force. We call this initial account of force Force[1]. Consider also that the student is guided through a Newtonian explanation of a puck sliding on ice at constant velocity, ignoring friction (i.e., the ice’s upward force against the puck counters the downward force of gravity on the puck, resulting in a zero net force). After consideration, she now confidently understands this phenomenon P1 with a Newtonian concept of force Force[2].

This raises several questions which have far-reaching implications. How is this new concept of force Force[2] stored relative to the old concept Force[1]? Does the Force[1] shift/change

directly, thus becoming Force[2]? Is an entirely new Force[2] concept added, e.g., as a placeholder or by copying and revising the old concept? We call this the problem of *information revision*. Figure 4 illustrates an extremely simplified topology of how information might be revised in the student's memory. The white nodes are phenomena in the student's memory, and the black nodes are categories (of force) that have been used to explain these phenomena. Figure 4(a) plots the comprehension of three phenomena P1-P3 before learning Force[2], and Figure 4(b-f) shows five possible accounts of the student's state after learning Force[2] and its relevance to P1. Figure 4 does not represent knowledge at the proper granularity for each of the four theories (e.g., for KiP, force is represented, in part, by a causal network), and it does not include all imaginable schemes of information revision. However, it is suitable for discussing differences in conceptual change theories. We discuss each of the information revision schemes shown in Figure 4, some assumptions behind them, and some implications for theories of conceptual change. We refer to the previously existing category (e.g., Force[1]) as the *prior* category, and the new/revised category (e.g., Force[2]) as the *subsequent* category.

If categories are directly revised as in Figure 4(b-c), then the prior category literally becomes the subsequent category, and afterward there is no trace of the prior. In the case of Figure 4(b), the learner immediately loses understanding of phenomena (e.g., P2 and P3) that were understandable in terms of the old concept but not in terms of the new concept. This seems unlikely, since students have access to their misconceptions after becoming acquainted with scientific concepts (Clement, 1982).



**Figure 4: Five possible accounts of how category information is revised. Black and white nodes represent categories and phenomena, respectively. Arrows indicate “is understood in terms of.” Dotted zones indicate contexts. (a) Initial state with category Force[1]. (b-f) Possible resultant states after incorporating Force[2].**

A second variety of direct revision is depicted in Figure 4(c): the prior category is directly revised into the subsequent category, and the learner immediately comprehends previous phenomena (e.g., P2 and P3) in terms of the subsequent category. This shares the same problem we mentioned for direct revision, and also creates more problems. First, it is unlikely that the Force[1] and Force[2] categories overlap perfectly in the range of phenomena they can explain, so a perfect substitution is not plausible. Additionally, there is empirical evidence that novices can utilize different conceptual knowledge based on the phenomena that needs to be explained. For instance, 70% of the novice subjects in diSessa et al. (2004) claimed that different forces were at work in phenomena similar to P2 and P3 described in Figure 4(a). Similarly, Collins and

Gentner (1987) interviewed a subject who explained two slightly different instances of evaporation with different, mutually incoherent, evaporation mechanisms.

Some of these problems with direct revision can be solved by assuming that prior and subsequent categories actually coexist for some time. In this case, conceptual change involves copying and revising (hereafter *copy-revising*) the prior knowledge to create a minimally or radically different subsequent knowledge. Copy-revision is shown in Figure 4(d), where the previous category is still used to understand P2 and P3, but understanding of P1 has been shifted to the subsequent concept. The prior knowledge (e.g., a substance-like category of force) and the subsequent knowledge (e.g., Newtonian force) have different existence conditions and consequences, so they are mutually incoherent. If we assume copy-revision happens, then we have many other questions to answer: How do people form coherent explanations with incoherent knowledge? How does a student eventually use the subsequent knowledge in place of the prior knowledge, where applicable? What mechanisms monitor the performance of the prior and subsequent concepts and shift their contexts?

A fourth possibility is shown in Figure 4(e): categories are copy-revised and the prior category is quarantined. In quarantine, the prior category cannot be used to explain new phenomena – it only exists until the phenomena it supports (e.g., P2 and P3) are understood in terms of other concepts. This makes the unlikely assumption that once a student has a small foothold in Newtonian dynamics she immediately discredits her prior intuitive concepts across all possible contexts.

A final information revision scheme is shown in Figure 4(f): categories are copy-revised and the subsequent Force[1] and consequent Force[2] categories are explicitly contextualized in Context[1] and Context[2], respectively. These contexts then behave as walls to maintain

internal coherence of the knowledge within. This solves the potential problem of incoherence and it allows the prior category Force[1] to continue to be utilized selectively. However, this also raises additional questions: Are new contexts established whenever any category revision occurs? What prevents a combinatorial explosion of contexts? How do phenomena (e.g. P1-P3) come to be understood in terms of the subsequent category in the new context?

None of the information revision schemes in Figure 4 are themselves complete theories of conceptual change. But they suggest that information revision – even at a very abstract level – has wide implications for theories of conceptual change, especially those that make claims about coherence and categorical shift.

For each of the four conceptual change theories, we discuss their commitment to how information is organized and revised with respect to a student learning Newtonian concept Force[2] in the presence of Force[1]. Some of the theories do not take a clear stand with respect to whether prior and subsequent concepts (i.e., beliefs, categories, and mental models) can exist simultaneously, so our analysis includes some speculation.

In Carey's (2009) account of Quinian bootstrapping, a student learning Newtonian dynamics would establish another conceptual system with a placeholder symbol for "force." This entails at least the following operations: (1) recognize that the present and new concepts of force are incoherent (i.e., incommensurable); (2) establish a new conceptual system  $CS_2$  for everyday dynamics; (3) create a placeholder symbol for the new force concept in  $CS_2$ ; (4) create placeholder symbols in  $CS_2$  for related concepts (e.g., acceleration and mass) and relations between them; and (5) enrich  $CS_2$  using modeling processes. These operations illustrate that Carey's theory does not involve direct revision of categories. Rather, it involves a very shallow copy-revision (more of an addition) since the subsequent concept is only a placeholder symbol.

This is most similar to Figure 4(f), where Context[1] represents CS<sub>1</sub> and Context[2] represents CS<sub>2</sub>, although both contexts are clearly lacking other quantities and placeholder symbols.

Coherence is enforced at the granularity of conceptual systems, since the prior and subsequent concepts are stored in different conceptual systems. Step 5 describes how the new conceptual system obtains content, but it is not clear how real-world phenomena come to be explained in terms of the new conceptual system CS<sub>2</sub> with Force[2] rather than the previous system CS<sub>1</sub> with Force[1].

Chi's (2008; Reiner et al., 2000) account of categorical shift is less straightforward with respect to the retention of previous beliefs and categories. The conjecture of Chi and colleagues is that the concept of force starts as a subordinate category of "Substances" for most novices, and then is shifted to become a subordinate of the lateral category "Constraint-based interactions" under the "Processes" ontological tree. Unlike Carey's theory, Chi's theory does not mention the establishment of a new conceptual system that permits Force[1] and Force[2] to coexist.

Ioannides and Vosniadou (2002) note that "Chi and colleagues seem to believe that conceptual change is a radical process that happens in a short period of time as an individual learns the correct ontology for a given concept" (pp. 7). In defense of Chi and colleagues, Chi (2008) notes that conceptual change only happens quickly if the learner is already familiar with the target category (e.g., "Constraint-based interactions") of the categorical shift. Otherwise, the learner must learn the properties of the target category, e.g., via formal instruction, before they can complete the categorical shift (Chi, 2008; 2005). So, Chi's theory of conceptual change is prolonged over the enrichment of the target category. After this is achieved, the concept Force[1] appears to be directly revised/shifted (e.g., as in Figure 4b-c), so the prior and subsequent concepts do not exist simultaneously. Further, this suggests that conceptual change

of the force concept would be trivial (or even instantaneous) if the learner was already familiar with “Constraint-based interactions.”

According to Vosniadou’s theory, changing the meaning of force is a gradual transition from an “initial” meaning of force through a series of “synthetic” meanings of force that incorporate aspects of the initial theory with scientific knowledge (Ioannides and Vosniadou, 2002). The overall change from intuitive to scientific concepts of force is gradual due to smaller changes in the beliefs and presuppositions (described above) that comprise the learner’s framework theory. Some of these changes in the meaning of force occur spontaneously. For example, a student with an *internal* meaning of force (i.e., force is an internal property of physical objects affected by weight and/or size) might notice that objects appear to *acquire* forces which sustain their movement. This is inconsistent with the idea that forces are only internal. Since the learner is committed to coherence, “*acquired* and *internal* force cannot coexist” (Ioannides & Vosniadou, pp. 41, their emphasis). Thus, the learner spontaneously shifts to an *acquired* meaning of force (i.e., objects acquire forces which cause movement).

The assertion that internal and acquired meanings of force cannot coexist suggests that Vosniadou’s theory involves directly revising the prior concept – or at least immediately eliminating it. Thus, in Vosniadou’s theory, the prior and subsequent concepts do not exist simultaneously. Had the authors stated that these meanings of force cannot coexist *in the same framework*, then we would conclude that Vosniadou’s mechanism of change involves quarantined copy-revision. Unlike Chi’s theory, Vosniadou’s theory segments the larger change from initial to Newtonian force into a series of incremental conceptual changes; however, like Chi’s theory, the individual changes are conducted by directly revising the framework theory and concepts embedded therein.



In the knowledge in pieces literature, diSessa and Sherin (1998) use the term *coordination class* to describe a connected set of strategies for gathering information and understanding the world. In this account, physical quantities (e.g., force and velocity) are considered coordination classes rather than categories (e.g., bird or hammer). This is because quantities often connect preconditions to particular outcomes in a *causal net* which is part of a coordination class. diSessa and Sherin use the equation  $F = ma$  to exemplify a causal net<sup>13</sup> since the existence of a force “causes” acceleration: we can determine force by observing acceleration and we can predict acceleration by knowing force. The authors perform an in-depth analysis on the interview transcript of a student describing the forces that exist when a hand pushes a book along the surface of a table. The authors explain the student’s problem-solving difficulties in terms of competing causal nets: a Newtonian  $F = ma$  causal net applies to the situation but makes predictions that she believes are inconsistent, so she excludes the situation from  $F = ma$  and instead uses an intuitive causal net. This suggests that intuitive and instructional conceptual structures – which are mutually incoherent – simultaneously coexist and compete to explain phenomena. This is a clear example of addition/copy-revision in Figure 4(d), where Force[1] and Force[2] indicate different coordination classes.

Our analysis suggests that there are disagreements among these theories on the foundational issue of how information is revised. Carey’s theory and KiP both involve the establishment of new conceptual structures that coexist with prior structures; however, the theories disagree on how the new and old structures are contextualized. Chi’s and Vosniadou’s theories apparently rely on the direct revision of concepts once the appropriate category of a concept is learned (according to Chi) or once the presuppositions and theories of the framework permit it (according to Vosniadou).

---

<sup>13</sup> Not all causal nets are equations, since students have many qualitative assumptions about quantities and causality.

One objection to this analysis is that theories of conceptual change theories can be noncommittal about how information is revised – after all, it is often advantageous to discuss cognition at different levels of abstraction (e.g., Marr, 1982). In counter-argument, each of these theories of conceptual change makes a claim about the presence or absence of coherence. Coherence has implications for the information revision scheme, and visa-versa. Consequently, conceptual change theories should describe the relationship between prior and subsequent knowledge, including whether they coexist and how they are contextualized.

The issue of whether new information coexists with previous, conflicting knowledge – and how it does so – has implications for coherence, the role of context, the mechanisms and complexity of change, and the process of understanding. I believe that most of the disagreements among conceptual change theories stem from vagueness and disagreement on this fundamental issue.

### **2.2.3 The effect of explaining on the process of change**

The research of Chi and colleagues (Chi et al., 1994a; Chi, 2000; de Leeuw & Chi, 2002) has characterized the *self-explanation effect*, where explaining new information to oneself helps repair flawed mental models. Chi et al. (1994a) determined that students who explain to themselves while reading a textbook passage - even when prompted by an experimenter to do so – perform better on a posttest than students who simply read the passage twice. Frequent self-explainers experience the greatest benefit. Chi (2000) describes the mechanism by which self-explaining promotes mental model transformation: (1) explaining the new knowledge causes recognition of qualitative conflicts (i.e., different predictions and structure) between a mental model and the text model; (2) the conflict is propagated in the mental model to find

inconsistencies in the consequences; and (3) the mental model is repaired using elementary addition, deletion, concatenation, or feature generalization operators. In short, self-explanation finds contradictions within implicit conflicts, thus causing belief revision. This can result in mental model transformation in Chi's framework, as described above.

Constructing an explanation for peer interaction can have the same beneficial effects on learning as self-explanation (Webb, 1989). Both explanation scenarios require that we make sense of relevant information; however, explaining to somebody else requires that we monitor the listener's comprehension, which might distract from our learning.

In Vosniadou's theory of conceptual change, "specific explanations" (synonymous with "specific theory;" Ioannides and Vosniadou, 2002) are embedded within a larger framework theory. It is not clear whether "specific explanation" refers to Chi's notion of explanation, but it appears that explanations – or the specific theoretical components thereof – are persistent structures (unlike Chi's theory). As in Chi's theory, constructing a new explanation can revise or replace these structures within the larger framework. Since we have too little information on how explanation affects conceptual change in Vosniadou's theory, we do not speculate any further.

At the heart of Carey's (2009) account of Quinian bootstrapping are modeling processes that provide meaning for placeholder structures in a new conceptual system. These modeling processes include analogy, induction, thought experiments, limiting case analyses, and abduction (i.e., reasoning to the best explanation). Both analogy and abduction are relevant mechanisms of explanation for our discussion.<sup>14</sup> These explanation processes generate the actual content of a

---

<sup>14</sup> Chi et al. (1994a) use the spontaneous analogy "the septum [of the heart] is like a wall" as an example of a self-explanation (pp. 454-455), so we include analogy in our discussion of the effect of explanation.

new conceptual system by (1) importing knowledge from other domains via analogy, and (2) making coherent assumptions via abduction.

Chi and Carey are assuming the same explanatory mechanisms (i.e., model-based abduction and analogy) but in reference to different types of change. Chi discusses how explanation promotes mental model transformation by repairing conflicts, and Carey discusses how it enriches a new conceptual system for incommensurable conceptual change. We believe that constructing explanations can play both of these roles, and our computational model constructs explanations to achieve both of these types of conceptual change (i.e., mental model transformation and category revision). Our computational model does not simulate all of the modeling processes mentioned by Carey (2009), nor does it model Quinian bootstrapping in its entirety.

From the KiP perspective, constructing an explanation involves combining and jointly using multiple pieces of knowledge. diSessa (1993) notes that using multiple p-prims in dynamic sequence or standard clusters accounts for these p-prims to raise or lower their *structured priority* simultaneously, where structured priority refers to (1) the strength of the connections between a p-prim and previously activated elements and (2) its likelihood of remaining activated during subsequent processing. This indicates that explaining shifts the context of conceptual structures. This, too, is a role of explanation in our computational model.

We see no explicit disagreement regarding the role of explanation in conceptual change. Each theory describes a separate effect of explaining, but these effects are mutually consistent.

### 2.2.4 The source of coherence

There is wide consensus that coherence is a desirable property of explanations (Thagard, 2007; Lombrozo, 2011), and that people revise their explanations to cohere with credible knowledge (Sherin et al., 2012). There is less agreement, however, on the source of coherence, and even on the definition of coherence (diSessa et al., 2004; Ioannides and Vosniadou, 2002; Thagard, 2000). Where the definition of coherence is more subjective, we discuss the dispute over the more general – and less ambiguous – epistemic property of logical consistency. In short, if a set of beliefs and mental models do not directly entail a contradiction, they are logically consistent.<sup>15</sup> Logical consistency is necessary but not sufficient for coherence. We do not assume that all possible contradictions are immediately detected by the learner, so for our discussion, “consistency” refers to perceived consistency rather than objective logical consistency. We discuss the disagreement among conceptual change theories about the role and source of consistency, which helps illustrate the more complicated dispute about coherence.

To begin, we must define coherence and consistency as a quantified property. A set of beliefs and mental models can be *internally consistent* if they do not entail a contradiction, regardless of beliefs and mental models outside of the set. Beliefs are *globally consistent* if the superset of all beliefs and models of the learner do not entail a contradiction. Internal and global coherence can be bounded in a similar fashion, but coherence is stricter than logical consistency.

Carey’s theory assumes coherence – and therefore logical consistency – within conceptual systems. When a learner utilizes a coherent, intuitive conceptual system  $CS_1$  and encounters an instructional concept that is incommensurable with  $CS_1$ , he or she establishes a new, coherent

---

<sup>15</sup> We do not assume that the set of beliefs and models is deductively closed, since this is not presumed of any of the theories of conceptual change. Consequently, we are referring to contradictions that are entailed directly from this knowledge.

conceptual system  $CS_2$ . While the learner acquires content and relation structure for  $CS_2$ , the knowledge in  $CS_1$  is still available. Conceptual systems  $CS_1$  and  $CS_2$  are internally consistent, but  $CS_1$  and  $CS_2$  may be mutually inconsistent, so the learner's knowledge may be globally inconsistent. For Carey, the granularity of consistency is at the level of conceptual systems, and it appears to be a hard constraint. Interestingly, a learner's knowledge *must* be globally incoherent in Carey's theory, since incoherence is a necessary property of incommensurability, and incommensurability is a precursor for establishing the new conceptual system  $CS_2$ . Consequently, Carey assumes internal coherence of conceptual systems and global incoherence among the union of all conceptual systems.

In Chi's theory, beliefs and mental models are revised when logical inconsistencies are detected. This is triggered via belief-level refutation or via self-explanation, which propagates implicit conflicts into explicit contradictions (Chi, 2008). In Chi's theory, consistency does not appear to be a hard constraint on conceptual systems, but the lack of consistency in a conceptual system drives the revision of components. Consistency therefore is a soft constraint (i.e., it is desired but not required).

In Vosniadou's theory, two inconsistent concepts (e.g., meanings of force) cannot coexist within the same framework theory (Ioannides and Vosniadou, 2002). When an inconsistency is detected within a framework theory, it is immediately remedied. This is because mental models are "dynamic, situated, and constantly changing representations that adapt to contextual variables" (Vosniadou, 2007, pp. 11). Unlike Carey's theory, Vosniadou's theory does not mention the establishment of a new context to store the inconsistent concept, so it is not clear whether the old concept exists. Since framework theories are internally consistent and inconsistent concepts are removed from them, Vosniadou's theory appears to assume global

consistency in a student's knowledge. However, Vosniadou (2007) further argues that students lack metaconceptual awareness of their beliefs, and that promoting this awareness is an integral part of teaching for conceptual change. This suggests that inconsistency and incoherence may frequently go undetected by novice students, which weakens this global coherence constraint considerably.

Knowledge in pieces involves the coexistence of new and old conceptual structures that are globally incoherent and that make globally inconsistent predictions. Coherence and consistency are therefore not properties of the knowledge system, but they are generally properties of the explanations that are constructed from it. When individual knowledge elements (e.g., p-prims) are combined to form a coherent explanation, their structured priorities are modified (diSessa, 1993). As a result, knowledge elements that are coordinated coherently (and therefore, consistently) are more likely to be activated together in the future. Coherence and consistency spread as new combinations of knowledge are considered and as knowledge elements are used in new contexts. Since the explanation process has a bias toward coherence, coherence emerges from this process rather than from the knowledge system directly.

In summary, there are direct disagreements about the source of consistency and coherence in explanations and knowledge systems. From the KiP perspective, the knowledge system is incoherent, and coherence is a product of coordinating knowledge into explanations based on dynamic activation priorities. In contrast, the other three theories rely on one or more generally coherent conceptual systems prior to explanation construction. According to Chi and Vosniadou, incoherence is a cue to modify a conceptual system by revising beliefs and mental models and the categories used to represent them. Carey agrees that incoherence can lead to belief revision and enrichment within a single conceptual system, but disagrees that it causes incommensurable

changes such as categorical shift within a single conceptual system. For Carey, when inconsistency is accompanied by incommensurability during formal education, it is a cue for establishing a new conceptual system altogether, which is internally coherent and consistent.

### **2.3 The path forward**

Our computational model of conceptual change can shed light on the areas of disagreement and divergence discussed in this chapter: how information is revised, the role of explanation, and the source of coherence. Our computational model is not an implementation of any of these four theories; the psychological assumptions of our model conflicts in some ways with each of the theories described above. Further, our model of conceptual change is not complete with respect to any of these theories – there are many things it does not model, including the following: (1) the development of metacognitive awareness of one’s beliefs (Vosniadou, 2007); (2) the full spectrum of model-based processes that enrich a new conceptual system (Carey, 2009); and (3) spontaneous analogies for self-explanation (Chi, 1994a). We therefore cannot expect this – or any – single cognitive model to reconcile all four theories outlined in this chapter. More reasonable goals for our computational model include the following: (1) develop a system for representing and contextualizing conceptual knowledge; (2) integrate the roles of explanation in each conceptual change theory into a single framework; and (3) demonstrate that a knowledge system can indeed be globally incoherent yet still produce coherent explanations.



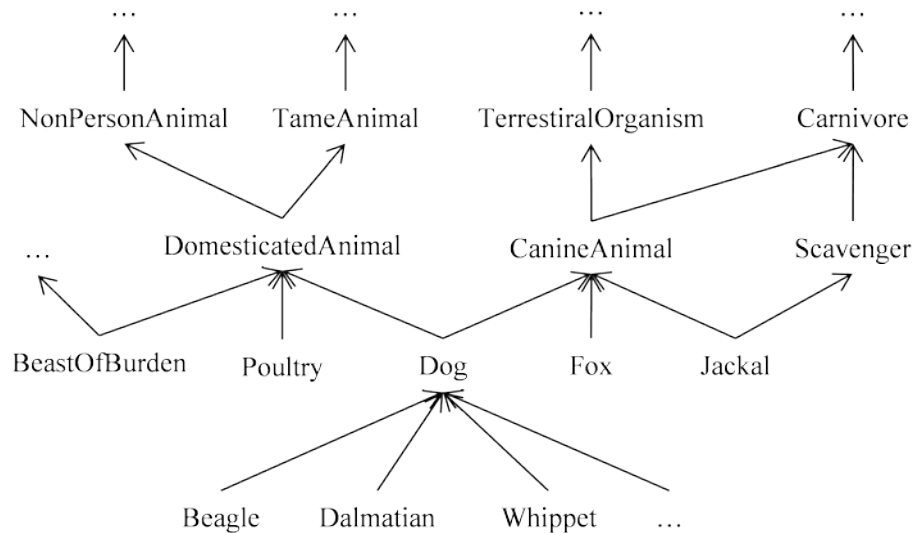
## Chapter 3: Background

Our computational model of conceptual change draws upon a number of areas of AI. For instance, qualitative modeling - a research area initially motivated by the study of human mental models - provides us with a composable knowledge representation and a vocabulary for descriptive and mechanism-based models. Computational cognitive models of analogical mapping, reminding, and generalization can be used for comparison, retrieval, and induction, respectively. We can also use existing AI technology for logically contextualizing information and for tracking the rationale of beliefs and their underlying assumptions. Finally, we can use existing tools to automatically encode sketches into relational knowledge, to rapidly and reliably create data for learning and testing in modalities familiar to people.

### 3.1 Ontologies

An *ontology* represents a set of categories (also called *collections*) and the relationships between them. Each category represents some type of object/substance (e.g., `Dog`, `ContainedFluid`, `HeartValve`) or event/situation (e.g., `FluidFlow`, `PhysicalTransfer`, `BuyingADrink`). These collections are part of the vocabulary with which beliefs are represented. For instance, we can assert the statement `(isa entity2034 Dog)` to say that the symbol `entity2034` is an instance of the collection `Dog`, or more casually, that `entity2034` is a `Dog`. Ontologies contain relationships between collections. For example, the statement `(genls Dog CanineAnimal)` states that all instances of the *subordinate* collection `Dog` are also instances of *superordinate* category `CanineAnimal`, but not necessarily the other way around. This makes ontologies hierarchical. Figure 5 illustrates a small portion of the OpenCyc ontology which includes `Dog`

and `CanineAnimal` collections. We use the OpenCyc ontology for our cognitive model, but we only use very small portions of it. The OpenCyc ontology was not constructed with the intent of modeling novice learners – quite the opposite, in fact – so we make heavy use the `isa` and `genls` relations but only minimal use of the abstract content.



**Figure 5: A small portion of the OpenCyc ontology. An arrow  $a \rightarrow b$  indicates `(genls a b)`.**

On a related note, Chi’s (2008) theory of conceptual change, outlined in Chapter 2, assumes the existence of “ontological trees.” These share the hierarchical property of the ontologies described here; however, it is not clear that categories in Chi’s ontological trees can inherit from multiple superordinate categories as illustrated in Figure 5.

### 3.2 Qualitative reasoning

In the introduction, we mentioned popular examples of conceptual change, including the changing concepts of force, heat, and temperature. Changes in other concepts such as speed, velocity, momentum, acceleration, mass, weight, light, and electricity have also been characterized in the literature (Reif, 1985; Dykstra et al., 1992; Reiner et al., 2000).

Interestingly, all of these concepts are represented as *quantities* at some point in the trajectory of misconceptions, and most of them are represented as quantities throughout. Consequently, modeling conceptual change involves representing and reasoning with quantities and also revising the existential and behavioral properties of quantities.

“Quantity” is not synonymous with “number.” A quantity (e.g., the volume of lemonade in a pitcher) may be assigned a numerical, unit-specific value (e.g. 12 fluid ounces) at a specific time. But people can reason very effectively without numbers. For instance, we might know that the volume of lemonade in a pitcher is greater than zero ounces and less than the volume of the pitcher (e.g., 64 ounces). If the height of the lemonade is millimeters below the rim of the pitcher, we might estimate that the volume is roughly *six-glasses-worth*, or just use a qualitative label such as *a lot* to represent the volume, based on how the estimate anchors within our space of experiences (Paritosh, 2004). Without numerical knowledge, we can also reason about causality. For example (quantities in italics), we know that if we increase the *angle of the pitcher*, the *height of the pitcher lip* will decrease. Once it decreases below the *height of the lemonade*, a fluid flow will start, and as we continue to increase the *angle of the pitcher*, we will also increase the *rate of flow*. In this example, we used the words “increase” and “decrease” to refer to the direction of change of a quantity’s value, and we used *ordinal relationships* such as “below,” to refer to inequalities between the values of two quantities. In this manner, people can reason qualitatively about continuous quantities, rates and directionalities of change, and ordinal relationships (i.e., greater than, less than, equal to) between them. A large literature describes formal approaches for representing and reasoning about processes (e.g., Forbus, 1984) and devices (e.g., de Kleer & Brown, 1984), and simulating systems provided this knowledge (e.g., Kuipers, 1986).

Novices and experts alike often reason with incomplete and imprecise qualitative knowledge, especially in situations of informational uncertainty (Trickett & Trafton, 2007). Consider the following incorrect *near-far* novice explanation of how the seasons change (Sherin et al., 2012): the earth orbits the sun along an elliptical path and is closer to the sun in the summer than in the winter. This mental model includes no numbers, but mentions quantities (e.g., the distance between the earth and the sun, the temperature of the earth) and relations between quantities (e.g., the earth's temperature strictly increases as its distance to the sun decreases). This is textbook qualitative reasoning. We next review relevant AI methods for representing, constructing, and reasoning with qualitative models.

### 3.2.1 Qualitative Process Theory

Qualitative process (QP) theory (Forbus, 1984) provides a vocabulary for representing mechanisms of change. Under QP theory, only *processes* cause changes in a physical system. For our example of pouring lemonade in the previous section, model fragments can represent the contained fluids and the flow of fluid.

QP theory also includes causal relationships between quantities. *Direct influences* are relationships between quantities where a quantity (e.g., the rate of flow) increases or decreases another (e.g., the volume of the fluid in the source). Direct influences often exist between the rate of a process and an affected quantity, and are represented by  $i+$  and  $i-$  relations (e.g., consequences of `FluidFlow` in Figure 6), which describe positive and negative direct influences, respectively. *Indirect influences* describe causal relationships between quantities where a quantity (e.g., the volume of a container) causes a positive or negative change in another quantity (e.g., the pressure of the fluid therein) under a closed-world assumption. Indirect

influences are represented by `qprop` and `qprop-` relations (e.g., consequences of `ContainedFluid` in Figure 6), which describe positive and negative indirect influences, respectively. Qualitative proportionalities represent causal influences between quantities where the direction of change is strictly increasing or decreasing.

```
ModelFragment ContainedFluid
Participants:
  ?con Container (containerOf)
  ?sub StuffType (substanceOf)
Constraints:
  (physicallyContains ?con ?sub)
Conditions:
  (greaterThan (Amount ?sub ?con) Zero)
Consequences:
  (qprop- (Pressure ?self) (Volume ?con))
```

When a container *con* physically contains a type of substance *sub*, a contained fluid exists. When there is a positive amount of *sub* in *con*, the volume of *con* negatively influences the pressure of this contained fluid.

```
ModelFragment FluidFlow
Participants:
  ?source-con Container (outOf-Container)
  ?sink-con Container (into-Container)
  ?source ContainedFluid (fromLocation)
  ?sink ContainedFluid (toLocation)
  ?path Path-Generic (along-Path)
  ?sub StuffType (substanceOf)
Constraints:
  (substanceOf ?source ?sub)
  (substanceOf ?sink ?sub)
  (containerOf ?source ?source-con)
  (containerOf ?sink ?sink-con)
  (permitsFlow ?path ?sub
    ?source-con ?sink-con)
Conditions:
  (unobstructedPath ?path)
  (greaterThan (Pressure ?source)
    (Pressure ?sink)))
Consequences:
  (greaterThan (Rate ?self) Zero)
  (i- (Volume ?source) (Rate ?self))
  (i+ (Volume ?sink) (Rate ?self))
```

When two contained fluids – a *source* and a *sink* – are connected by a *path*, and both are of the same type of substance, a fluid flow exists. When the *path* is unobstructed and the pressure of *source* is greater than the pressure of *sink*, the rate of the flow is positive and it decreases the volume of *source* and increases the volume of *sink*.

**Figure 6: `ContainedFluid` (above) and `FluidFlow` (below) model fragments used in the simulation in Chapter 7. English interpretations for the model fragments included at right.**

### 3.2.2 Compositional modeling

In compositional modeling (Falkenhainer & Forbus, 1991), domain knowledge is represented using *model fragments*, which are combinable pieces of domain knowledge. Modeling the flow

of blood in the circulatory system (see Chapter 7 for detail) involves a number of model fragments, two of which are shown in Figure 6: the conceptual model fragment `ContainedFluid`, and the process model fragment `FluidFlow`. Model fragments are *instantiated* during reasoning. For example, we might infer `ContainedFluid` *model fragment instances* when reasoning about the human circulatory system since each of the chambers of the heart contain blood. Each model fragment  $m$  can be uniquely defined by a tuple  $\langle P, C, A, N, S \rangle$ , which includes participants, constraints, assumptions conditions, and consequences, respectively. We describe these using the model fragments in Figure 6 as an example.

*Participant statements* ( $P$ ) are statements describing the entities involved in the phenomenon. For example, the `?con` participant in `ContainedFluid`, is of type `Container`, so for the entity `heart` to fill the `?con` participant role, it must be a `Container`, so the statement `(isa heart Container)` must be true for `heart` to bind to `?con`. Each participant statement is a statement such as `(isa ?con Container)` which states that the participant slot (e.g., `?con`) must be of a specific type (e.g., `Container`). Participant slot `?con` also has relational role `containerOf`, so `(containerOf cf heart)` would be true of any `ContainedFluid` instance `cf` where `heart` is bound to `?con`.

*Constraints* ( $C$ ) are statements that must hold over the participants in order for an instance of the model fragment to exist. When the constraints hold, an instance *instance*( $m, P$ ) of model fragment  $m$  is inferred as a distinct entity over the participants  $P$ . For example, if `(physicallyContains heart Blood)` is true of `Container` instance `heart` and `StuffType` instance `Blood`, then a new model fragment will be instantiated with participant bindings  $B = \{ \langle ?con, heart \rangle, \langle ?sub, Blood \rangle \}$ . Logically, model fragment instantiation can be

expressed as the following first-order logical equivalence, where a conjunction of two sets of statements is the conjunction of the union of member statements:

$$P \wedge C \equiv \mathbf{instance}(m, B).$$

*Modeling assumptions* ( $A$ ) are statements concerning the model fragment's relevance to the task at hand. These make the granularity, perspectives, and approximations of the model fragment explicit. These help select the appropriate method of description for problem solving, since the behavior of a single physical phenomenon (e.g., blood flow through arteries) can be described at multiple granularities (e.g., describing fluid volumes or describing localized collections of matter being transported through the body). Our computational model does not use modeling assumptions to simulate students, but we do believe that students are capable of reasoning at different levels of description, and that learning the appropriate level of description for problem-solving is important for achieving expert understanding. This is future work.

*Conditions* ( $N$ ) are propositions that must hold over a model fragment's participants that limit the model fragment's behavioral scope, such as `(greaterThan (Amount ?sub ?con) Zero)` in `ContainedFluid`. Conditions differ semantically from constraints, since an instance of a model fragment can exist without a condition satisfied. When all conditions of a model fragment instance hold, the instance is *active*. More formally:

$$\mathbf{instance}(m, B) \wedge A \wedge N \equiv \mathbf{active}(\mathbf{instance}(m, B)).$$

*Consequences* ( $S$ ) are propositions that describe a model fragment instance's constraints on a system's behavior when it is active. For example, the unground consequence

```
(qprop- (Pressure ?self) (Volume ?con))
```

of `ContainedFluid` is inferred as

```
(qprop- (Pressure ch) (Volume heart))
```

when an instance `ch` is active with participant bindings  $B = \{\langle ?con, heart \rangle, \langle ?sub, Blood \rangle\}$ .

This imposes the constraint that the pressure of the contained fluid `ch` increases as the volume of `heart` decreases. Model fragment activation can be expressed as the following logical implication:

$$\mathbf{active(instance}(m, B)) \rightarrow S.$$

Inference with model fragments can therefore be summarized with the implication

$$P \wedge C \wedge A \wedge N \rightarrow S.$$

Model fragments are instantiated and activated within a *scenario*, which is a logical context that contains a partial description of the phenomena to be modeled, such as the propositional facts and rules about the solar system for using the model fragments in Figure 6. Model fragments are stored within a *domain theory*, which is a set of model fragments and scenario-



independent beliefs. The result of model formulation is a *scenario model* composed of one or more model fragment instances. Importantly, one model fragment instance may serve as a participant of another (e.g., FluidFlow in Figure 6 has two ContainedFluid participants: ?source and ?sink), so the resulting scenario model may have a nested structure.

Provided compositional models and qualitative process theory, what constitutes a “concept” in our model of conceptual change? Put simply, a concept is domain knowledge that can be learned and revised. We define the three following types of knowledge as concepts:

- **Model fragments:** The model fragments in Figure 6 and others (e.g., interaction of forces, floating, sinking, fluid flow, and heat flow) represent concepts because they are learnable (see Chapter 5) and revisable (see Chapter 8). As mentioned in Chapter 1, model fragments represent parts of human mental models.
- **Categories and quantities:** Chapter 8 describes how the quantities within compositional model fragments can be ontologically revised using heuristics, so quantities such as force, heat, and sunlight are also concepts.
- **Propositional beliefs:** Domain-level propositional beliefs about the world are concepts, according to the common phrase “the concept that  $p$ ” where  $p$  is a proposition such as “the earth orbits the sun.” The truth value of these propositions can change in our model. We do not consider metaknowledge propositions (e.g., the proposition that I learned about the aorta from a textbook) to be concepts.

The term “concept” has obvious problems due to its ambiguity, so we refer to the specific components – model fragments, quantities, and propositional beliefs – when possible, and

compactly use the term “conceptual knowledge” or “concept” to refer to all three types of knowledge at once.

We must also define the term “misconception” in the context of our model. In the literature, misconceptions are often stated as general propositions such as, “continuing motion implies a continued force in the direction of the movement” (Chi, 2005). In our model, misconceptions are mistakes produced by a theory comprised of model fragments, beliefs, and quantities. For example, in Simulation 1, the qualitative models learned by the system produce the misconceptions that (1) surfaces do not push up against objects resting on their surface and (2) objects pushed in a given direction always go in that direction, irrespective of prior velocity. These misconceptions are exhibited on specific scenarios, but we can conclude that the system would perform similarly on analogous scenarios due to the principles of model-based inference described above.

### 3.3 Abductive reasoning

*Abduction* can be defined as reasoning to the best explanation for a set of observations (Peirce, 1958). In AI, this has been formalized as a search for some set of assumptions<sup>16</sup> that can prove the observations,<sup>17</sup> where an *explanation* for the observations is a set of assumptions and justification structure that together infer the observations. This amounts to searching for the best set of assumptions that explain the observations. Abduction has been used in AI for plan recognition, diagnosis, language interpretation, and other tasks.

Systems that use abduction must at least computationally implement a *better* comparator between explanations so that they can search for the best explanation. Depending on the task,

---

<sup>16</sup> Assumptions are also referred to as *hypotheses* in the AI abduction literature.

<sup>17</sup> Observations are also referred to as *evidence* in the AI abduction literature.

explanatory preference might rely on which explanation is more probable (e.g., Pearl, 1988), which makes fewer assumptions (e.g., Ng & Mooney, 1992), or which makes less costly assumptions (e.g., Charniak & Shimony, 1990; Santos, 1994). Cost-based abduction (CBA) is of particular relevance to this dissertation, where the goal is to find a least-cost proof (LCP) where each assumption has a weighted cost. Finding LCPs is NP-Hard (Charniak & Shimony, 1994), and so is approximating LCPs within a fixed ratio of the optimal solution (Abdelbar, 2004).

Our model of conceptual change uses abductive reasoning to construct explanations for new and previously-encountered observations. We describe our abductive reasoning algorithm in Chapter 4, but it is worth pointing out similarities with existing approaches here. A more accurate term for our explanation construction process is *abductive model formulation* since our model uses qualitative model fragments to represent domain knowledge and composes them into a scenario model via model formulation, described above. The explanation evaluation process – whereby the agent determines the best explanation – is similar to CBA, but differs in two important ways to model humans: (1) consistency is a soft constraint (i.e., contradictions are permitted but costly) within and across explanations; and (2) more than just assumptions have a cost, e.g., model fragments, model fragment instances, contradictions, and other elements. In CBA, individual assumptions have weighted costs, but in our model, some sets of beliefs (e.g., those comprising a logical contradiction) also have costs.

### **3.4 Analogical processing**

Two simulations described in this thesis utilize analogical reasoning. This involves matching the relations and entities among two cases to make similarity judgments, generalizations, and inferences. We briefly review these analogical subsystems next.

### 3.4.1 The Structure-Mapping Engine

The Structure-Mapping Engine (SME) (Falkenhainer et al., 1989) is a domain-general computational model of analogy and similarity, based on Gentner's (1983) structure-mapping theory of analogy. Its inputs are two cases, the *base* and *target*, consisting of structured representational statements. SME computes one or more mappings between the base and the target. Each mapping contains (1) *correspondences* that match expressions and entities in the base with expressions and entities in the target, (2) a numerical *structural evaluation score* of the quality of the mapping, and (3) *candidate inferences* that assert what might hold<sup>18</sup> in the target. Candidate inferences may not be deductively valid, but they may produce useful hypotheses (e.g., Gentner, 1989; McLure et al., 2010; Christie & Gentner, 2010). We will refer to the following functions of SME in the below:

- ***best-mapping(*b*, *t*)***: returns the SME mapping with the highest structural evaluation score, using base *b* and target *t* cases as input.

The SME structural evaluation score can be normalized by dividing it by the maximum self-score, (i.e., the maximum score attained by matching either the base or target to itself). This ensures that  $0 \leq \text{normalized score} \leq 1$ . We use the following functions to refer to structural evaluation scores:

- ***sim-score(*m*)***: returns the numerical structural evaluation score of a SME mapping *m*.

---

<sup>18</sup> Since they are the product of structural similarity alone, candidate inferences are not necessarily deductively valid; however, they are useful hypotheses (e.g., Gentner, 1989; McLure et al., 2010; Christie & Gentner, 2010).

- ***self-score(c)***: returns the numerical structural evaluation score of a SME mapping between a case  $c$  and itself. Computed as ***sim-score(best-mapping(c, c))***.
- ***norm-score(m)***: returns a normalized structural evaluation score  $s$ , such that  $0 \leq s \leq 1$ , for SME mapping  $m$  with base  $m.base$  and target  $m.target$ . Computed as:

$$\frac{\text{sim-score}(m)}{\max(\text{self-score}(m.base), \text{self-score}(m.target))}$$

### 3.4.2 MAC/FAC

MAC/FAC (Forbus et al., 1995) is a domain-general computational model of similarity-based retrieval. Its inputs are (1) a *probe* case and (2) a *case library* (set of cases). Cases consist of structured, relational statements, like the inputs to SME. MAC/FAC retrieves one or more cases from the case library that are similar to the probe via a two-stage filtering process. The first stage is coarse, using a vector representation automatically computed from the cases to estimate similarity between the probe and the contents of the case library by computing dot products in parallel. It returns the case library case with the highest dot product, plus up to two others, if sufficiently close. The second stage uses SME to compare the probe with the cases returned by the first stage. It returns the case with the highest similarity score, plus up to two others, if sufficiently close. The mappings it computes are available for subsequent processing. We use the following functions to describe MAC/FAC retrieval:

- ***macfac(p, C)***: given a probe case  $p$  and a case library  $C$ , returns an ordered sequence  $M$  of mappings retrieved via MAC/FAC, where  $0 \leq |M| \leq 3$ . Sequence  $M$  is ordered such that ***sim-score(m<sub>i</sub>) ≥ sim-score(m<sub>i+1</sub>)***, so the most similar MAC/FAC retrieval is  $m_0$ , and the most similar case is  $m_0.target$ .

- ***macfac-best***( $p, C$ ): returns the first element (highest-similarity mapping) of ***macfac***( $p, C$ ).

### 3.4.3 SAGE

SAGE (Friedman et al., in preparation) is a computational model of analogical generalization that uses both SME and MAC/FAC. SAGE clusters similar examples into probabilistic generalizations, where each generalization typically describes a different higher-order relational structure. SAGE takes a sequence of positive examples  $E = \langle e_0, \dots, e_n \rangle$  represented as cases, and a numerical *similarity threshold*  $s$  ( $0 \leq s \leq 1$ ) as its inputs. SAGE produces (1) a set of *generalizations*  $G = \{g_0, \dots, g_i\}$ , each of which is a probabilistic case created by merging similar examples in  $E$ , and (2) a set of *ungeneralized examples*  $U = \{u_0, \dots, u_j\} \subseteq E$ , that were not sufficiently similar to other examples to generalize.

SAGE is initialized with  $G = U = \emptyset$ . When given a new example  $e_i \in E$ , SAGE calls ***macfac-best***( $e_i, G \cup U$ ) to find the best mapping  $m$  between  $e_i$  and an existing generalization or ungeneralized example. If there is no such mapping or the mapping is below the similarity threshold (i.e., ***norm-score***( $m$ )  $< s$ ) then the new example is added to the list of ungeneralized exemplars (i.e.,  $U = U + e_i$ ) and the algorithm terminates. Otherwise, SAGE merges  $e_i$  and the case that was retrieved via MAC/FAC. The merge happens differently depending on whether MAC/FAC retrieved an ungeneralized example or a generalization. If the retrieved case is an ungeneralized example  $u$  then (1) a new generalization  $g$  is created by merging  $e_i$  with  $u$ , (2) the size of  $g$  is set to two (i.e.,  $|g| = 2$ ), (3)  $g$  is added to  $G$ , and (4) the  $u$  is removed from  $U$ . If MAC/FAC retrieved an existing generalization  $g$ , then  $e_i$  is merged into  $g$ , and the size of  $g$  is incremented by 1 (i.e.,  $|g| = |g| + 1$ ).

When SAGE merges a new example  $e$  with a previous case  $c$  (i.e., a previous example or generalization), it records a probability for each statement to represent its frequency within the resulting generalization (Halstead & Forbus, 2005). The probability of a statement  $s$  within the resulting generalization  $g$  is a factor of (1) the probabilities of  $s$  in  $e$  and  $c$  and (2) the size of  $c$ , written as  $|c|$ . If  $c$  is an ungeneralized example,  $|c|$  equals 1; otherwise,  $|c|$  is the number of cases that has been merged into the generalization  $c$ . We compute the probability of any statement  $s$  in the resulting generalization  $g$  as follows:

$$P(s \text{ in } g) = \frac{P(s \text{ in } c)|c| + P(s \text{ in } e)}{|c| + 1}.$$

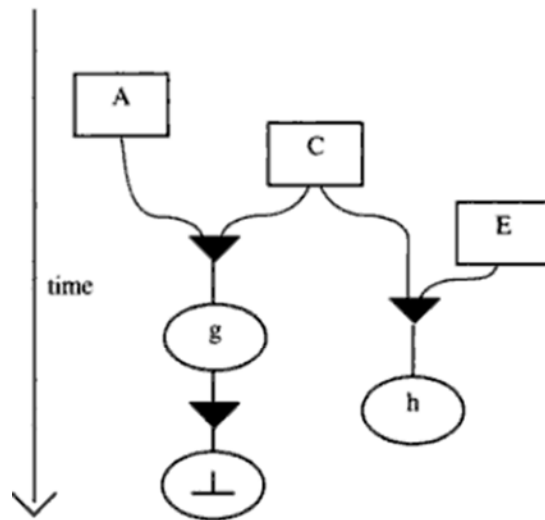
Note that any statement  $s$  not present in  $e$  or  $c$  has probability  $P(s \text{ in } e) = 0$  and  $P(s \text{ in } c) = 0$ , respectively. For the case where two examples are merged into a new generalization, all statements with correspondences in the mapping are inferred with a probability of 1.0, and all expressions without correspondences in the mapping are inferred with a probability of 0.5.

Using this merge technique, common relational structure is preserved with a probability of 1.0, and non-overlapping structure is still recorded, but with a lower probability. The probability affects similarity judgments in SAGE. This is because the individual similarity score of each SME match hypothesis is weighted by the probability of the corresponding base and target statements. Consequently, low-probability expressions in a generalization contribute less to similarity judgments.

SAGE has been used for concept learning in domains such as sketch recognition, spatial prepositions, and clustering short stories (all three in Friedman et al., 2011), as well as for learning sentence structure from example sentences describing events (Taylor et al., 2011).

### 3.5 Truth Maintenance Systems

A *Truth Maintenance System* (TMS) communicates with an inference engine to track the justifications for beliefs (Forbus & de Kleer, 1993). Tracking the justifications for beliefs improves problem-solving efficiency in three ways relevant to our conceptual change model:



**Figure 7: A TMS containing assumptions (squares), justified beliefs (ovals), justifications (triangles), and a contradiction  $\perp$  node (courtesy Forbus & de Kleer, 1993)**

1. Explanations can be generated via a justification trace.
2. The system can identify the faulty foundations – including assumptions – of a bad conclusion.
3. *Caching* inferences by retaining them in justification structure is generally more efficient than re-running the inference process all over again.<sup>19</sup>

<sup>19</sup> If inference rules are few and inexpensive to run, caching inferences may actually degrade performance.



Specialized types of TMSs exist, but our model of conceptual change uses a JTMS (justification-based TMS), so we only review the details relevant to JTMSs. For our purposes, a TMS includes a network of *belief nodes* that represent distinct beliefs and *justifications* which associate zero or more *antecedent* belief nodes with a *consequent* belief node. There are different types of belief nodes, three of which are shown in the example TMS in Figure 7:

1. A *premise* node represents a belief that holds universally.
2. An *assumption* node represents a belief that can be explicitly enabled (believed) or retracted (disbelieved) by the agent.
3. A normal belief node represents a belief that is believable iff it is justified by other beliefs.
4. A *contradiction* represents a logical inconsistency within the justifying beliefs. For example, in Figure 7, belief node *g* supports a contradiction, which is supported by assumptions *A* and *C*, so at least one of *A* and *C* is faulty. For the sake of conserving existing beliefs, assumption *A* may be retracted to avoid retracting support for *h*.

In a TMS, multiple justifications can justify a single belief node. This indicates that the belief has more than one unique line of reasoning for believing it. Suppose we want to find an explanation for a belief in the TMS for abductive reasoning. Explanations for a belief node *n* in a TMS are based on *well-founded support* (Forbus & de Kleer, 1993) for that node. Well-founded support is any sequence of justifications  $J_1 \dots J_k$  such that:

- Node *n* is justified by  $J_k$ .

- All antecedents of  $J_k$  are justified earlier in the sequence.
- No belief node has more than one justification in the sequence.

In Figure 7,  $h$  has well-founded support from its supporting justification, provided assumptions  $C$  and  $E$  are enabled. The contradiction has well-founded support from its supporting justification and the justification supporting  $g$ , provided  $A$  and  $C$  are enabled. If  $A$  is retracted, both the contradiction and  $g$  will lose all well-founded support. In this thesis, we call each set of possible well-founded support a *well-founded explanation*. Importantly, when a belief  $n$  is justified by two beliefs, it has *at least* two well-founded explanations, and it may have an exponential number of them.

TMS justification structure is used within our conceptual change model to track the rationale for beliefs. The definition of well-founded explanations dictates how the justification structure is aggregated into different explanations in our model. We discuss this further in Chapter 4.

### 3.6 Microtheory contextualization

Conceptual learning at the scale we advocate in this thesis requires a large knowledge base (KB) – both quantitatively, in the number of different facts, and qualitatively, in the number of different predicates and entities. As the knowledge base grows, storing all propositional beliefs, rules, and mental models in a single logical context would quickly make reasoning intractable. In many learning systems, the control knowledge that initially speeds up learning and reasoning eventually degrades performance. This has been called the *utility problem* (Minton, 1990).

Aside from tractability issues, conceptual change involves reasoning with competing, potentially inconsistent knowledge. This requires the use multiple logical contexts.<sup>20</sup> Representing inconsistent explanations requires representing inconsistent beliefs, and when this occurs within the same logical context, it entails a contradiction. A contradiction within a logical context entails any belief via indirect proof – for AI systems, but not necessarily for people – which is problematic for reasoning about the state of the world.

Intractability can be mitigated and inconsistency can be tolerated by contextualizing the KB into hierarchical logical contexts that we call *microtheories*. Microtheories are hierarchical because a microtheory  $m_{child}$  can inherit from another microtheory  $m_{parent}$ , so that all statements in  $m_{parent}$  are visible in  $m_{child}$ . This allows us to quickly define logical contexts for reasoning without copying propositional beliefs. Contextualizing large KBs is not a new idea – there exist algorithms for automatically creating KB partitions (e.g., Amir & McIlraith, 2005) and for performing model formulation in a microtheory-contextualized KB (Forbus, 2010).

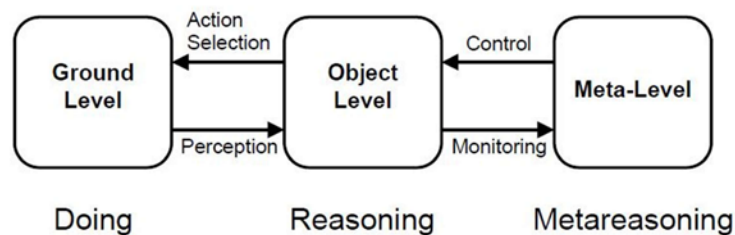
In the system described below, each microtheory in the KB contains zero or more relational statements, and each relational statement in the KB belongs to one or more microtheories. Microtheories are ubiquitous in the system described here: explanations are represented, in part, by microtheories; SME cases and SAGE generalizations are microtheories; model formulation uses microtheories for scenario descriptions, scenario models, and domain theories; and the entire explanation-based network described below is encoded as relational statements across several microtheories.

---

<sup>20</sup> Temporal and logical qualification predicates, (e.g., OpenCyc’s binary *holdsIn* relation) can be used to contextualize propositional beliefs within the same logical context so as to avoid entailing a contradiction; however, this is not necessarily the case for contextualizing rules, plans, and model fragments.

### 3.7 Metareasoning

As discussed above, reasoning with conceptual knowledge produces explanations about the world. But the process of conceptual change requires reasoning *about* the conceptual knowledge and *about* the explanations produced, to determine which beliefs are more productive and which explanations better suit the observations. We can therefore draw a distinction between (1) *object-level* reasoning with domain knowledge and (2) *meta-level* reasoning about object-level reasoning. Figure 8 illustrates both control and monitoring from the meta-level. In AI, *metareasoning* is the deliberation over plans and strategies available to an agent, and then selecting a course of action (Horvitz, 1988; Russell & Wefald, 1991; Cox, 2005). Since metareasoning can observe object-level operations, it can also be used for explaining these operations (e.g., Kennedy, 2008) and doing introspective learning (e.g., Leake & Wilson, 2008).



**Figure 8: Meta-level control and monitoring (Cox & Raja, 2007)**

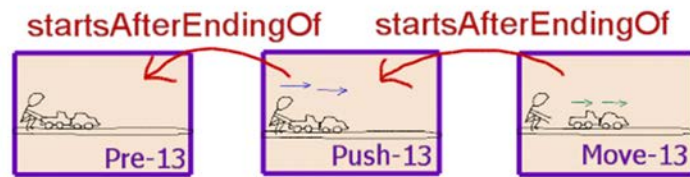
In our model of conceptual change, meta-level *monitoring* tasks include evaluating explanations (the product of object-level reasoning) and detecting anomalies within observations. Meta-level *control* tasks include (1) heuristic-based revision of knowledge and (2) preference encoding over concepts and explanations, both of which influence future object-level reasoning.

Knowledge can also be encoded at the meta-level. In our computational model, this includes knowledge *about* domain knowledge, such as (1) an explicit preference for one meaning

of *force* over another, (2) knowledge that the anatomical concept `LeftVentricle` was learned from a textbook, (3) knowledge that two explanations for the changing of the seasons are in competition, and so-forth. This *metaknowledge* aids in making decisions for future learning and reasoning.

### 3.8 CogSketch

CogSketch (Forbus et al., 2008) is an open-domain sketching system. CogSketch interprets the ink drawn by the user, and computes spatial and positional relations (e.g., `above`, `rightOf`, `touches`) between objects. Further, CogSketch supports multiple *subsketches* within a single sketch. We use this feature to create *comic graphs* (e.g., Figure 9) that serve as stimuli, where each subsketch in a stimulus represents a different qualitative state, and transitions between them



**Figure 9: A comic graph stimulus created using CogSketch.**

represent state changes. Similar stimuli have been used in analogical learning experiments with people (e.g., Chen, 1995; 2002).

Figure 9 depicts a stimulus from the simulation in Chapter 5. Each subsketch represents a change in the physical system illustrated. Within each subsketch, CogSketch automatically encodes qualitative spatial relationships between the entities depicted, using positional and topological relationships. For example, the person in Figure 9 is `above` and `touching` the ground in all three states, but the person and the toy truck are not `touching` in the third state.

Physical quantities such as area and axis coordinates are also computed by CogSketch and stored using relations and scalar quantities. For example, the statement

```
(positionAlongAxis truck-4 Horizontal (Inches 220))
```

asserts that entity `truck-4` is 220 inches to the right of the origin along the `Horizontal` axis.

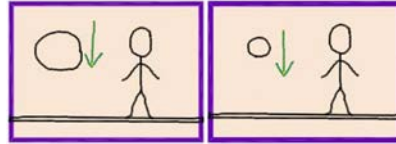
The arrows within a subsketch (e.g., the blue and green arrows in Figure 9) are user-generated *annotations* between objects, which represent relationships such as applied force (blue arrows) and movement (green arrows). The arrows between subsketches indicate temporal order, via the `startsAfterEndingOf` relation. Using quantity data, annotations, and temporal relations, the system can identify changes in physical quantities across states, which we refer to as *physical behaviors*. CogSketch is used to encode physical behaviors comprising the training and testing data for two of the simulations presented below. Since CogSketch automatically encodes the knowledge for these simulations, the knowledge representation choices were not made with the learning task in mind, so the stimuli were not hand-tailored for the specific learning tasks.

### 3.8.1 Psychological assumptions about using comic graphs

Although it is not background material per se, it is fitting to discuss the psychological assumptions we make by using sketched comic graphs as testing and training data. We begin by describing how the simulations in this dissertation use sketches for testing and training.

Experimenters in cognitive psychology and learning science frequently use multi-state sketches (like Figure 9) to describe a phenomenon occurring over time and then ask the subject for predictions or explanations (e.g., Hestenes et al., 1992; Chen, 1995; 2002). Other

experimenters use sketches such as Figure 10 and ask the subject to compare two scenarios (Ioannides & Vosniadou, 2002; diSessa et al., 2004). We refer to these *sketched testing data*. The simulations in Chapter 5 and Chapter 8 use the same sketched testing data as experimenters, redrawn by hand in CogSketch to be automatically encoded into relational knowledge for use by



**Figure 10: A sketch with two subsketches, redrawn from diSessa et al. (2004).**

the simulation. Using sketched testing data with CogSketch makes several assumptions about how people encode sketched knowledge, which we discuss below.

The simulations in this dissertation also use sketches for learning. For example, the sketch in Figure 9 is used by the simulation in Chapter 5 to learn humanlike preconceptions of pushing, moving, and blocking. This use of *sketched training data* is very different from sketched testing data. We list five considerations that arise from our choice of using comic graphs as learning stimuli:

1. *Real-life observations are represented as independent comic graph episodes.* As inhabitants of a continuous world, people must learn when a *jumping* event starts and ends, rather than being told the relevant start and end state in a comic graph. Since we provide the system with clear-cut cases such as Figure 9, we do not expose the system to distracting qualitative states that might occur before or after the event.

2. *Observations in a continuous world are approximated by a sequence of still pictures.*

The simulations are not observing a world of continuous – and continuously changing –

physical quantities. Instead, they are given CogSketch's output: qualitative spatial relations over objects and numerical values of spatial quantities. The sketched data therefore conveys relative changes in position, but not relative changes in velocity, so the simulation does not need to differentiate velocity from acceleration, which is difficult for novice students (Dykstra et al., 1992).

3. *The sequence of events is already segmented into different qualitative states.* The simulations do not have to find the often-fuzzy boundaries between physical behaviors as an event unfolds over time. In the Figure 9 example, the person pushes the truck, then the truck and car move, and then the truck and car stop – there is no temporal ambiguity in this chain of events.
4. *The objects and events in the stimuli are relevant to the concept being learned.* This is a factor of the sparseness of the sketches – they contain few confusing events, e.g., a dozen birds flying overhead, a broken wheel on a toy truck, and so forth. As a result, there are less confounds for inferring causality between events.
5. *The encoding in the stimuli is relevant to the concept being learned.* The encoding is sparse in that CogSketch does not encode knowledge about the internal components individual glyphs, e.g., that the head of the person in Figure 9 is an oval with a major axis angle of 39 degrees. Consequently, the qualitative relations produced by CogSketch comprise the majority of the encoding, and these are especially relevant for learning a qualitative theory of dynamics (e.g., Chapters 5 and 8). The output of CogSketch is not nearly as rich as human visual perception; however, we do believe that CogSketch captures an important subset spatial knowledge that people encode. This is not to say that the sketches contain no extraneous data; they contain entity attributes (e.g., *Truck*)



and extraneous conditions (e.g., the truck in Figure 9 is touching the car while it moves, which is not a necessary condition for movement) that must be factored out by learning algorithms.

All of these consequences of our sketched-based approximation of the real world are reasons to expect our simulations to learn real-world concepts much faster than people. Despite the differences between comic graphs and the real world, we believe that using automatically-generated training data is a significant advance over using hand-coded stimuli to simulate real-world experiences. We discuss more specific implications of these representation choices in the simulation chapters, where relevant.

## Chapter 4: A Computational Model of Explanation-Based Conceptual Change

This chapter describes our computational model of conceptual change. Except for the AI technologies discussed in Chapter 3, the computational model described in this chapter is a novel contribution of this dissertation. We describe how knowledge is contextualized using explanations and how constructing and evaluating explanations affects the knowledge of the agent. This provides the explanatory power of the system, and is especially relevant to the third claim of this dissertation:

***Claim 3:*** Human mental model transformation and category revision can both be modeled by iteratively (1) constructing explanations and (2) using meta-level reasoning to select among competing explanations and revise domain knowledge.

The core of our model includes the following: (1) a network for organizing knowledge; (2) an abductive algorithm for constructing explanations in the network; (3) meta-level strategies for selecting a preferred explanation; and (4) strategies for retrospectively explaining previously-encountered phenomena. This core model satisfies Claim 3.

After we describe how knowledge is organized, we describe the specifics of how explanations are constructed, retrieved, and reused. We then describe how preferences are computed over explanations, which drives the adoption and propagation of new information.

### 4.1 Two micro-examples of conceptual change

We consider the following two micro-examples of conceptual change in the remainder of this chapter:

1. **Circulatory system example** (mental model transformation, from Chapter 7): The agent's mental model of the circulatory system involves a loop from a single-chamber model of the heart to the body and back. After incorporating knowledge from a textbook, the agent revises its mental model so that (1) the heart is divided into left and right sides and (2) blood flows to the body from the left side of the heart.
2. **Force example** (category revision, from Chapter 8): The agent uses a force-like quantity  $q$  that is present in all objects. The agent cannot explain why a small ball travels farther than a large ball when struck by the same foot using its present concept of  $q$ . Consequently, the agent revises  $q$  so that it is transferrable between colliding objects, where the amount transferred is qualitatively inversely proportional to the size of the struck object.

These two examples are not isolated changes; they are part of larger model transformations (e.g., that the blood from the body flows to the right side of the heart and is then pumped to the lungs) and trajectories of change (e.g., that forces exist between, and not within, objects) in their respective simulations. But for ease of explanation in this chapter, here we consider them in isolation. Both types of conceptual change use the same core explanation-based framework described here, but category revision requires some additional operations. For instance, the category revision simulation in Chapter 8 uses heuristics to revise a quantity in the domain theory. We discuss operations specific to category revision in Chapter 8.

## 4.2 Contextualizing knowledge for conceptual change

Conceptual change involves managing inconsistent knowledge. The agent must encode beliefs and models that are inconsistent with prior knowledge, use them to reason about the world, and then determine which of the available beliefs and models provide the best (i.e., simplest, most accurate, and most credible) account. As we discussed in Chapter 3, we can divide knowledge into logical microtheories to retain local consistency where it's important. Our model uses microtheories (1) as sets of beliefs and model fragments and (2) as cases for analogical reasoning. We begin by discussing how microtheories are used to contextualize different types of information.

Recall the following from our compositional modeling discussion in Chapter 3: (1) a *scenario* is a set of statements that describes a problem; (2) the *domain theory* is a set of scenario-independent model fragments and statements; and (3) *model formulation* is the process of constructing a model of the scenario from elements of the domain theory. It is important for the agent to have record of what information was gathered from an external scenario (e.g., via observation or reading) and what was inferred via model formulation. This is achieved by representing each scenario as its own *scenario microtheory*.<sup>21</sup> In the circulatory system micro-example, multiple scenario microtheories contain the information from the textbook, and in the force example, two scenario microtheories contain the information about two observations: a foot kicking a small ball; and the same foot kicking a large ball. Each scenario microtheory is annotated with metaknowledge (defined in Chapter 3) that records the source of the information (e.g., observation, textbook, or interaction with another individual).

---

<sup>21</sup> See section 3.6 for a discussion of microtheories.

Some beliefs in a scenario microtheory describe processes, states, and events that the agent must explain. Following Hempel and Oppenheim (1948), we call these *explanandums*. Consider the circulatory system micro-example above: the agent encounters information from a textbook that the (1) heart is divided into two sides and (2) that blood is pumped from the left side to the body. This textbook-based scenario microtheory contains propositional beliefs describing objects such as

```
(isa l-heart (LeftRegionFn Heart))
```

which states the symbol `l-heart` is an instance of `(LeftRegionFn Heart)`. It also includes beliefs that together describe a single situation, such as

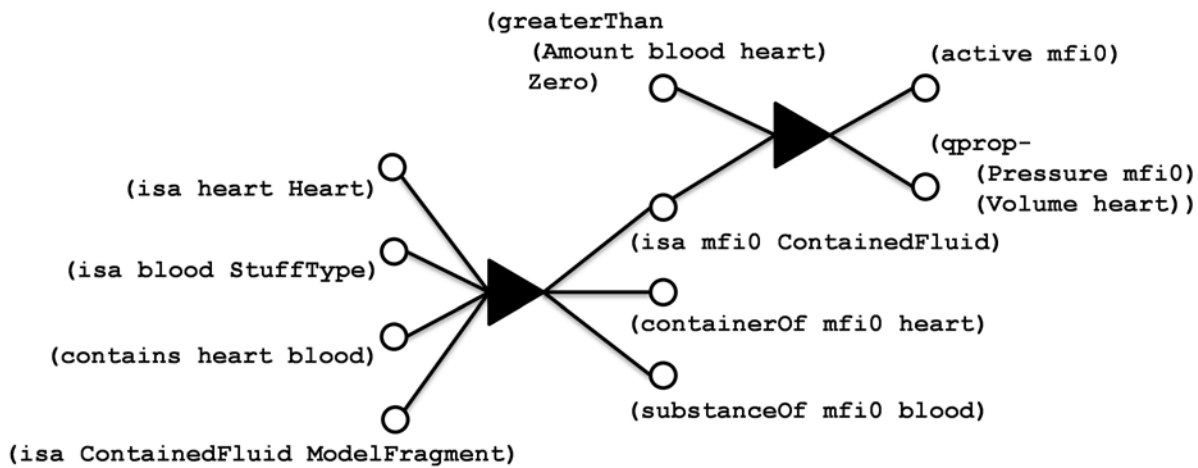
```
(isa leftH2B PhysicalTransfer)
(outOf-Container leftH2B l-heart)
(into-Container leftH2B body)
(substanceOf leftH2B Blood)
```

which describes `leftH2B`, the flow of blood from `l-heart` to the body. The four propositional beliefs describing the flow event `leftH2B` constitutes a single explanandum. When a new explanandum is encountered in a scenario, it is explained via model formulation.

When the agent encounters a new scenario such as the textbook information above, the scenario microtheory is added as a parent of the *domain knowledge microtheory*  $\mathbb{D}$ . Recall that when a microtheory is the parent of another, its statements are inherited by the child microtheory.  $\mathbb{D}$  thereby inherits all information from observations, interactions, and instruction

that the agent has encountered. In addition to inheriting from scenarios,  $\mathbb{D}$  also contains model fragments that have been induced from observations (e.g., via SAGE in Chapter 5). Importantly, information in one scenario microtheory may contradict information in another scenario microtheory, so the information in  $\mathbb{D}$  may be inconsistent (i.e., its conjunction could entail a logical contradiction). Propositional beliefs in  $\mathbb{D}$  may serve as premises.\*

When the agent constructs an explanation via model formulation, it uses subsets of  $\mathbb{D}$  as the domain theory and the scenario since  $\mathbb{D}$  inherits scenario information and contains model



**Figure 11:** A small portion of justification structure generated from model formulation in the circulatory system micro-example. The justification (triangle) at left is the logical instantiation of model fragment instance `mfi0` based on the constraints of `ContainedFluid` (see Figure 6 for `ContainedFluid` definition) and the justification at right is the logical activation of `mfi0`.

fragments. The output of model formulation includes (1) statements that are logically entailed by instantiating and activating model fragments, (2) assumptions\* that justify other beliefs, but have no justification themselves, and (3) justifications\* that associate antecedent and consequent statements. Figure 11 shows some justification structure resulting from model formulation in the circulatory system micro-example. Some belief nodes in Figure 11, e.g., `(contains heart blood)`, describe the specific structure of the circulatory system. These are in  $\mathbb{D}$  and inherited

\* This term is defined in section 3.5.

from scenario microtheories. The belief `(isa ContainedFluid ModelFragment)` in Figure 11 refers to the model fragment `ContainedFluid` which is also present in  $\mathbb{D}$ . Other belief nodes in Figure 11 (e.g., `(isa mfi0 ContainedFluid)`, `(containerOf mfi0 heart)`, and `(active mfi0)`) describe the scenario model. These beliefs are not visible from  $\mathbb{D}$ . They are stored in the *provisional belief microtheory*  $\mathbb{B}$  which contains beliefs generated via model formulation.

The distinction between  $\mathbb{D}$  and  $\mathbb{B}$  is that  $\mathbb{D}$  includes assertions about the world (e.g., `(contains heart blood)`: “*the heart contains blood*”) and models for reasoning about the world (e.g. `ContainedFluid`). In compositional modeling, you would find this information in scenarios and domain theories, respectively.  $\mathbb{B}$  contains the inferences (e.g., `(containerOf mfi0 heart)`: “*The container of the contained fluid mfi0 is the heart*”) and assumptions that result from reasoning with the information in  $\mathbb{D}$  and  $\mathbb{B}$ . Propositional beliefs in  $\mathbb{D}$  are believable (but not necessarily believed) independently of  $\mathbb{B}$ , but beliefs in  $\mathbb{B}$  use  $\mathbb{D}$  as a foundation for inference and assumption. This means that  $\mathbb{B}$  contains the scenario models produced by model formulation.

The rationale for each inference and assumption in  $\mathbb{B}$  is recorded using the justification structure produced via model formulation. We defined justifications in our discussion of truth maintenance systems in Chapter 3, but note that our justifications have multiple consequences.<sup>22</sup> The justification structure is recorded as propositional statements in a justification microtheory. For instance, the rightmost justification in Figure 11 is described by the following statements:

```
(isa j1 Justification)
```

---

<sup>22</sup> A justification with multiple consequences can be converted into a set of multiple justifications – one for each consequence – by creating a single-consequence justification with the same set of antecedents for each consequence.

```

(antecedentsOf j1 (greaterThan (Amount blood heart) zero))

(antecedentsOf j1 (isa mfi0 ContainedFluid))

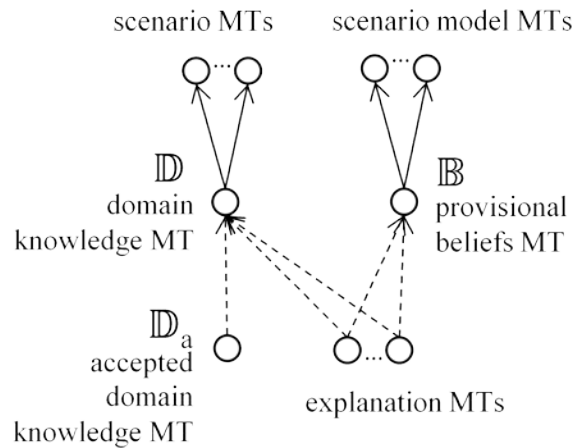
(consequencesOf j1 (active mfi0))

(consequencesOf j1 (qprop- (Pressure mfi0) (Volume heart)))

```

The justifications produced by model formulation are used to reify explanations and construct *explanation microtheories*. Each well-founded explanation in the justification structure corresponds to a different explanation, and the beliefs in each well-founded explanation are stored in separate explanation microtheories.

The final microtheory of note is the *adopted domain knowledge microtheory*  $\mathbb{D}_a$ . This is the subset of  $\mathbb{D}$  that the agent presently accepts as true. This does not mean that the agent explicitly



**Figure 12: The relationship between microtheories (MTs) in our computational model. Solid arrows represent “inherit all information from” (i.e., child-of), and dotted arrows represent “contains some information from.”**

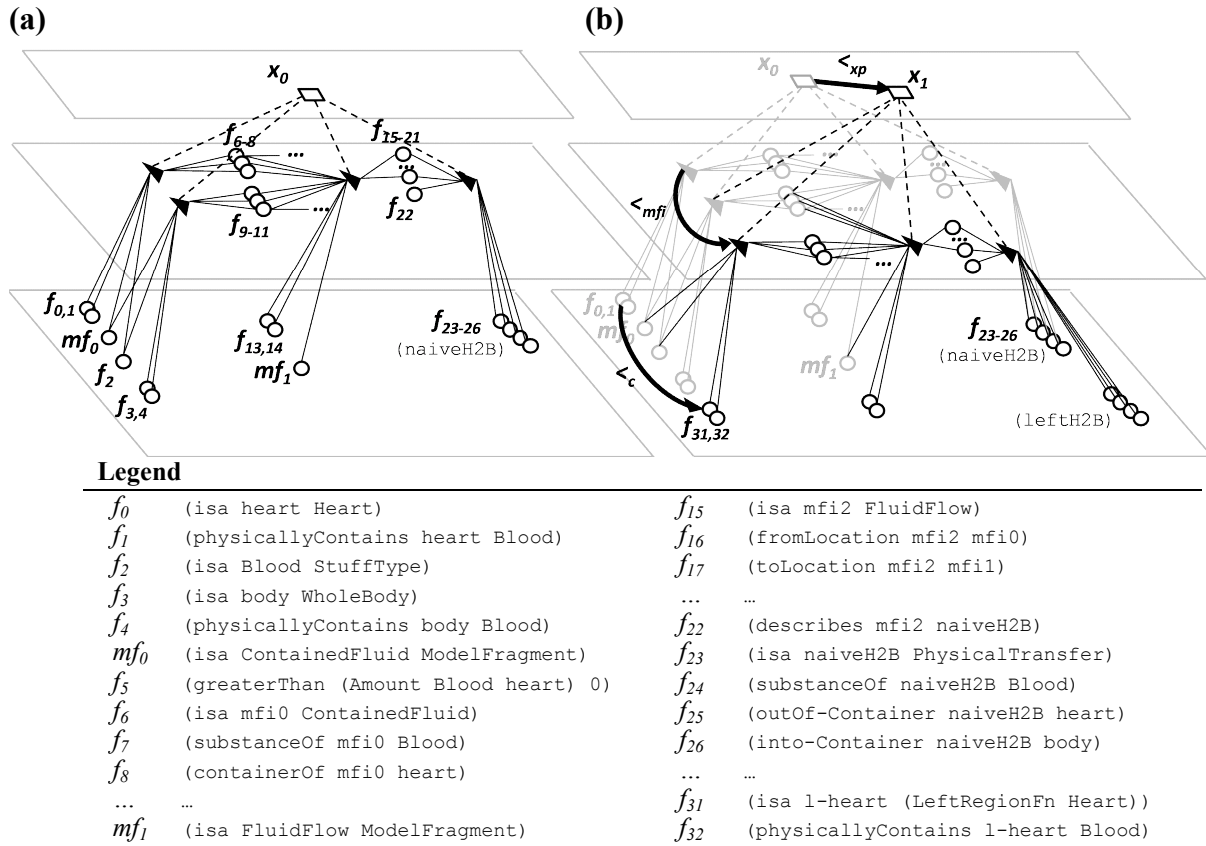
regards the beliefs in  $\mathbb{D}$  that are not present in  $\mathbb{D}_a$  (which we write  $\mathbb{D}/\mathbb{D}_a$ ) as false; rather, the agent may be undecided on the truth value of these beliefs. Like  $\mathbb{D}$ ,  $\mathbb{D}_a$  is not necessarily internally consistent. If  $\mathbb{D}_a$  is inconsistent, nothing is broken – we can simply say that the agent



holds beliefs to be true that are logically inconsistent.  $\mathbb{D}_a$  will become important later in this chapter, during our discussion of belief revision using cost functions.

The relationships between different microtheories and microtheory types that we have discussed are shown in Figure 12. The contexts  $\mathbb{D}$  and  $\mathbb{B}$  are collector microtheories of scenarios and scenario models, respectively. Explanation microtheories contain subsets of information from  $\mathbb{D}$  and  $\mathbb{B}$  that collectively participate in a well-founded explanation. Finally,  $\mathbb{D}_a$  contains the subset of the information from  $\mathbb{D}$  which is presently believed by the agent.

The remainder of our discussion of our computational model relies on this information organization scheme. We next describe how explanations, justifications, and beliefs are related. For quick reference, condensed definitions of the above microtheories and of other terms used later in this chapter are included in a table in the appendix.



**Figure 13: A portion of an explanation-based network. (a) Single explanation  $x_0$  for an explanandum  $naiveH2B$  (rightmost nodes). (b) After new knowledge is added, preferences are computed for new knowledge ( $<c$ ), new model fragment instances ( $<mfi$ ), and for the new explanation  $x_1$  ( $<xp$ ).**

### 4.3 An explanation-based network for conceptual change

Explanations, justifications, and beliefs can be viewed as a network that supports metareasoning and conceptual change. This is an extension of a justification structure network (e.g., Figure 11). A portion of a network is shown in Figure 13, before (Figure 13a) and after (Figure 13b) for the circulatory system micro-example outlined above. The legend of Figure 13 labels the key beliefs and model fragments for reference, but the specific beliefs are not yet important. We describe

the network with respect to this example. To improve readability, we lay out the network on three tiers. We describe them from bottom to top.

### **Bottom (domain knowledge) tier**

The bottom tier of the network in Figure 13(a-b) is the domain knowledge tier, and contains information from  $\mathbb{D}$ . This includes propositional beliefs, specifications of quantities, and model fragments. The bottom tier of Figure 13(a-b) contains the subset of  $\mathbb{D}$  that is relevant to the circulatory system micro-example. All propositional beliefs on this tier are supported by observation or instruction.

### **Middle (justification) tier**

The middle tier plots provisional beliefs from  $\mathbb{B}$  (represented as circles in Figure 13) and justifications (represented as triangles in Figure 13). As in Figure 11, the antecedents of a justification are on its left, and its consequences are on its right. The provisional beliefs and justifications in Figure 13(a-b) are the subsets that are relevant to the circulatory system micro-example. All of the justifications in the system are plotted on this tier. Unlike the bottom tier, the belief nodes on this tier are not supported by observation or instruction – they are inferred during the explanation construction process, which we describe in section 4.4.

### **Top (explanation) tier**

The top tier plots explanation nodes. Figure 13(a-b) depicts a subset of all explanations  $\mathbb{X}$  constructed by the agent, plotted with quadrilateral nodes  $x_0$  and  $x_l$  on the top tier. Each

explanation represents a well-founded explanation for some situation or belief. Each explanation is uniquely defined as  $x = \langle J, B, M \rangle$ , where

- $M$  is set of one or more explanandums  $M$  that are explained by  $x$ .
- $J$  is a set of justifications  $J$  that comprise a well-founded explanation (defined in Chapter 3) for  $M$ . In Figure 13, each explanation node has dashed lines to its justifications  $J$ .
- $B$  is the set of all beliefs that comprise the explanation.  $B$  includes all antecedents and consequences of the explanation's justifications  $J$ . This includes domain knowledge from  $\mathbb{D}$  and provisional beliefs from  $\mathbb{B}$ , so  $B \subseteq \mathbb{D} \cup \mathbb{B}$ . The explanation's microtheory contains all beliefs in  $B$ .

Based on these definitions, the network in Figure 13(a-b) tells us a lot about the agent's learning in the circulatory system micro-example. Before encountering the textbook information (Figure 13a), the agent justifies the flow of blood to the body `naiveH2B` with an explanation  $x_0$  that involves a `FluidFlow` process and two `ContainedFluid` instances: one for the heart and one for the rest of the body. There are no other explanations for this phenomenon. After the textbook scenario is incorporated (Figure 13b), the agent has information in  $\mathbb{D}$  about the left heart (`l-heart`) and the flow of blood from the left heart to the body (`leftH2B`). Figure 13b also contains a new, second explanation  $x_1$  which uses new and old information in  $\mathbb{D}$  (the bottom tier) and  $\mathbb{B}$  (the middle tier). The new explanation  $x_1$  justifies the old (`naiveH2B`) and new (`leftH2B`) situations, but note that the previous explanation  $x_0$  and its constituent justifications and beliefs still exist. These explanations are now in competition. In the following, we discuss

how explanations are constructed, how they compete, how they are reused, and how competitions are resolved to achieve conceptual change.

#### 4.4 Constructing explanations

Our computational model constructs explanations for an explanandum  $m$  in two steps: (1) perform abductive model formulation to create one or more scenario models that justify  $m$ ; (2) for each well-founded explanation of  $m$  within the resulting justification structure, create an explanation node in the network. Since computing well-founded explanations is described in Chapter 3, we concentrate here on our abductive model formulation algorithm which is a contribution of this research.

As stated above, compositional model fragments simulate parts of mental models. Figure 14 shows two model fragments: `ContainedFluid` and `FluidFlow`. Figure 13 contains the belief nodes  $mf_0$  (`isa ContainedFluid ModelFragment`) and  $mf_1$  (`isa FluidFlow ModelFragment`) which are used to explain blood flowing from the heart ( $x_0$ ) and the left-heart ( $x_1$ ) to the body. We use these explanations as examples for our description of explanation construction.

```

ModelFragment ContainedFluid
Participants:
  ?con Container (containerOf)
  ?sub StuffType (substanceOf)
Constraints:
  (physicallyContains ?con ?sub)
Conditions:
  (greaterThan (Amount ?sub ?con) Zero)
Consequences:
  (qprop- (Pressure ?self) (Volume ?con))

ModelFragment FluidFlow
Participants:
  ?source-con Container (outOf-Container)
  ?sink-con Container (into-Container)
  ?source ContainedFluid (fromLocation)
  ?sink ContainedFluid (toLocation)
  ?path Path-Generic (along-Path)
  ?sub StuffType (substanceOf)
Constraints:
  (substanceOf ?source ?sub)
  (substanceOf ?sink ?sub)
  (containerOf ?source ?source-con)
  (containerOf ?sink ?sink-con)
  (permitsFlow ?path ?sub
    ?source-con ?sink-con)
Conditions:
  (unobstructedPath ?path)
  (greaterThan (Pressure ?source)
    (Pressure ?sink))
Consequences:
  (greaterThan (Rate ?self) Zero)
  (i- (Volume ?source) (Rate ?self))
  (i+ (Volume ?sink) (Rate ?self))

```

When a container *con* physically contains a type of substance *sub*, a contained fluid exists. When there is a positive amount of *sub* in *con*, the volume of *con* negatively influences the pressure of this contained fluid.

When two contained fluids – a *source* and a *sink* – are connected by a *path*, and both are of the same type of substance, a fluid flow exists. When the *path* is unobstructed and the pressure of *source* is greater than the pressure of *sink*, the rate of the flow is positive and it decreases the volume of *source* and increases the volume of *sink*.

**Figure 14:** ContainedFluid (above) and FluidFlow (below) model fragments used in the simulation in Chapter 7. English interpretations of each model fragment (at right).

Before describing our algorithm, it is important to note that “abductive model formulation” is not synonymous with “abduction.” Abduction computes the set of assumptions (or hypotheses) that best explains a set of observations. In contrast, abductive model formulation computes qualitative models of a phenomenon by assuming the existence of entities and relations between them. If we later compare these qualitative models to compute the best explanation, then we have performed a nontraditional type of abduction. This section discusses the construction of the qualitative models, and we discuss the comparison of qualitative models later in this chapter.

Our abductive model formulation algorithm starts with the procedure *justify-explanandum*, shown in Figure 15, which is given three items as input:

1. A domain context  $D$  which is a microtheory that contains a subset of the model fragments in  $\mathbb{D}$ .
2. A scenario context  $S$  which is a microtheory that contains propositional beliefs (i.e., no model fragments).  $S$  contains a subset of domain knowledge in  $\mathbb{D}$ , since  $\mathbb{D}$  inherits from scenario microtheories which are necessary for model formulation.  $S$  also contains provisional beliefs from  $\mathbb{B}$  (from previous model formulation attempts) to reuse previous solutions. For example, if the agent has previously determined that there is a `ContainedFluid` instance within the heart, it need not reconstruct this.
3. An explanandum  $m$  that requires explanation. Our algorithm takes in two different types of explanandums: (1) propositional beliefs; and (2) entities that describe processes, e.g., `naiveH2B` which describes the transfer of blood from heart to body. When an explanandum is a belief, the algorithm directly justifies the belief, and when the explanandum is a process entity, the algorithm instantiates models that describe the entity. For our example, we will use the process entity `naiveH2B` as the explanandum, which is described by facts  $f_{23-26}$  in Figure 13.

Arguments  $S$  and  $D$  can be constructed from one or more explanations. For instance, using a set of explanations  $\{\langle J_0, B_0, M_0 \rangle, \dots, \langle J_n, B_n, M_n \rangle\}$ , we can construct  $S$  as a microtheory that contains all beliefs in the belief sets  $\{B_0, \dots, B_n\}$  of the explanations and we can construct  $D$  as

the set of all model fragments instantiated in these belief sets. We discuss this further in section 4.5 below.

When the explanandum provided to *justify-explanandum* is a process instance, the procedure *justify-process* does the rest of the work. Otherwise, when the explanandum is a proposition describing a quantity change or an ordinal relationship, the procedures *justify-quantity-change* and *justify-ordinal-relation*, respectively, do the rest of the work. To be sure, there are other types of propositions that can be justified, but since our simulations involve explaining processes and state changes, these explanandums and procedures are sufficient for the simulations in this thesis.



## Front-ends to abductive model formulation

```

procedure justify-explanandum(explanandum m, domain D, scenario S)
  if m is a symbol and m is an instance of collection C such that (isa C ModelFragment):
    justify-process(m, D, S)
  else if m unifies with (greaterThan ?x ?y):
    justify-ordinal(m, D, S)
  else if m unifies with (increasing ?x) or with (decreasing ?x):
    let q, d = quantity-of-change(m), direction-of-change(m)
    justify-quantity-change(q, d, D, S)

procedure justify-process (process instance m, domain D, scenario S)
  // Find collections of the given entity within D
  let C = query D for ?x: (isa m ?x)
  // Find model fragments in D that are specializations of these collections.
  let F = query D for ?x: c ∈ C ∧ (isa ?x ModelFragment) ∧ (genls ?x c)
  for each f in F:
    // Find participant roles {⟨slot0, role0⟩, ..., ⟨slotn, rolen⟩} of f
    let P = participant-roles-of(f)
    // Find entities in S that fill participant roles of a f instance describing m
    let R = query S for ⟨slot, ?x⟩ : ⟨slot, role⟩ ∈ P ∧ (role m ?x)
    abductive-mf-instantiation(f, R, D)

procedure justify-ordinal-relation (ordinal relation m, domain D, scenario S)
  // m is of the form (greaterThan (MeasurementOf q s1) (MeasurementOf q s2))
  let q, s1, s2 = quantity-of(m), state-1-of(m), state-2-of(m)
  if query S for (after s2 s1) then:
    justify-quantity-change(q, i-, D, S)
  if query S for (after s1 s2) then:
    justify-quantity-change(q, i+, D, S)

procedure justify-quantity-change (quantity q, direction d, domain D, scenario S)
  // Find direct and indirect influences of q
  instantiate-fragments-with-consequence(qprop q ?x), D, S)
  instantiate-fragments-with-consequence(qprop- q ?x), D, S)
  instantiate-fragments-with-consequence(d q ?x), D, S)
  let Ii = query S for indirect influences on q. // results are in form (qprop/qprop- q ?x)
  for each i in Ii:
    let di = direction-of-influence(i) // qprop or qprop-
    let qi = influencing-quantity(i)
    let dc = d
    if di = qprop- then:
      set dc = opposite(d)
    justify-quantity-change(qi, dc, D, S)

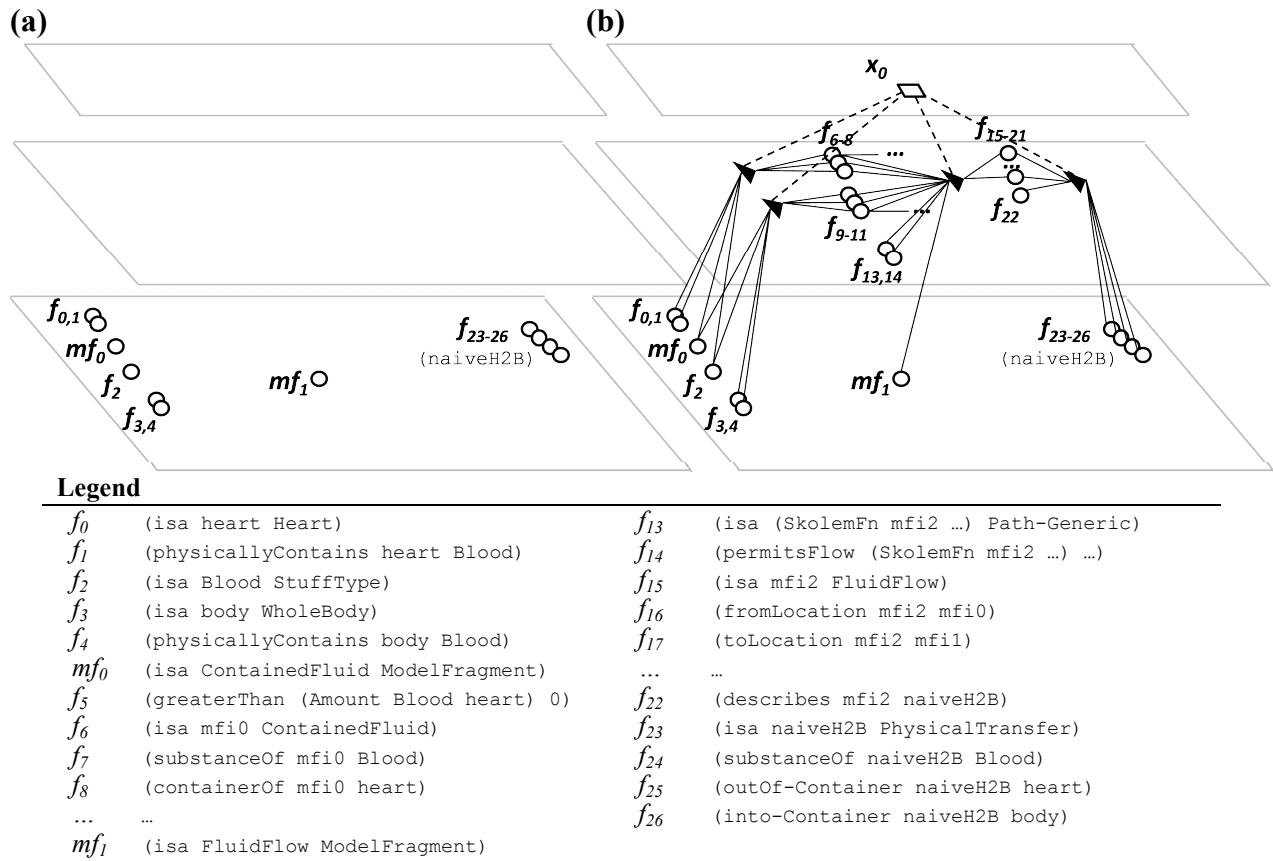
```

**Figure 15: Pseudo-code for front-end procedures that trigger abductive model formulation.**

Regardless the type of explanandum that is being explained, all paths through *justify-*

*explanandum* call the procedure *abductive-mf-instantiation*. This procedure takes a model fragment *m* (e.g., `FluidFlow`), a *role binding list* *R* that associates zero or more participant slots of the model fragment with known entities (e.g.,  $\{\langle ?\text{sub}, \text{Blood} \rangle, \langle ?\text{source-con}, \text{heart} \rangle, \langle ?\text{sink-con}, \text{body} \rangle\}$ ), and the *D* and *S* arguments from *justify-explanandum*. It instantiates and activates all possible instances of *m* that conform to the role binding list *R* provided the scenario information *S*. Importantly, if it cannot bind some participant slot to an entity within *S*, it will assume the existence of an entity that satisfies this role, and it will assume propositions (i.e., constraints and conditions) as necessary. For example, if there is no known `Path-Generic` instances that satisfies the constraints for `?path` participant of `FluidFlow`, the algorithm will assume the existence of such an entity.

We begin by stepping through an example of explanation construction that uses *justify-process*. The behaviors of the *justify-quantity-change* and *justify-ordinal-relation* procedures are discussed in Chapter 6. We use the explanation of situation `naiveH2B` in Figure 13 as an example.



**Figure 16: A portion of explanation-based network. (a) Before an explanation has been constructed for naiveH2B. (b) After an explanation  $x_0$  has been constructed for naiveH2B via abductive model formulation.**

Suppose that the agent's knowledge is in the state depicted in Figure 16(a): the agent believes, due to PhysicalTransfer instance naiveH2B, that there is a transfer of blood from the heart to the body. However, there is no knowledge of a path or specific process by which this occurs. When the agent explains naiveH2B with the call *justify-explanandum*(naiveH2B,  $D, S$ ), the procedure first determines whether naiveH2B can be justified by a model fragment. Since naiveH2B is a PhysicalTransfer, the system will check whether there are model fragments that can model a PhysicalTransfer. Suppose the belief (genls FluidFlow PhysicalTransfer) is present in  $D$ , indicating that this is indeed the case.

The next step is to find properties of `naiveH2B` that are important for modeling it as a `FluidFlow`. Consider the following participant roles of `FluidFlow` from Figure 14:

```
?source-con Container (outOf-Container)
?sink-con Container (into-Container)
?source ContainedFluid (fromLocation)
?sink ContainedFluid (toLocation)
?path Path-Generic (along-Path)
?sub StuffType (substanceOf)
```

The procedure must next search for participants for each of the following slots:  $\{?source-con, ?sink-con, ?source, ?sink, ?path, ?sub\}$ . If it cannot find a participant in the scenario, it must either instantiate a model to fill the role or assume the existence of the participant. We discuss each of these cases. First, some of these participants can be found in  $S$ . For example, the participants `?source-con`, `?sink-con`, and `?sub` correspond to the roles `outOf-Container`, `into-Container`, and `substanceOf`, respectively. The procedure queries  $S$  to determine which entities (if any) fill these roles of `naiveH2B`:

```
(outOf-UnderSpecifiedContainer naiveH2B ?source-con)
(into-UnderSpecifiedContainer naiveH2B ?sink-con)
(substanceOf naiveH2B ?sub)
(fromLocation naiveH2B ?source)
(toLocation naiveH2B ?sink)
(along-Path naiveH2B ?path)
```

Not all of this information is present in  $S$ , but some information about `naiveH2B` is represented as  $f_{24-26}$  in Figure 16:

```
(outOf-UnderSpecifiedContainer naiveH2B heart)
(into-UnderSpecifiedContainer naiveH2B body)
(substanceOf naiveH2B blood)
```

From these assertions, the procedure constructs the *binding list*  $R = \{\langle ?source-con, heart \rangle, \langle ?sink-con, body \rangle, \langle ?sub, Blood \rangle\}$  to bind the participant variables to *ground* (i.e., non-variable) entities in  $S$ . More work must be done: the three remaining participant slots (i.e., `?source`, `?sink`, and `?path`) must be bound and constraints must be tested in order to explain `naiveH2B` with a `FluidFlow` instance. This is handled by calling ***abductive-mf-instantiation***(`FluidFlow`,  $R$ ,  $S$ ,  $D$ ) in Figure 16.

Abductive instantiation of `FluidFlow` with partial bindings  $R$  begins by finding participants that are themselves model fragments. This includes `?source` and `?sink`, both of which are `ContainedFluid` instances. The procedure finds constraints on these `ContainedFluid` instances by substituting the bindings  $R = \{\langle ?source-con, heart \rangle, \langle ?sink-con, body \rangle, \langle ?sub, Blood \rangle\}$  into the `FluidFlow` constraints. This substitution produces the following set of statements:

```
(substanceOf ?source Blood)
(substanceOf ?sink Blood)
(containerOf ?source heart)
```

```
(containerOf ?sink body)
(permitsFlow ?path Blood heart body)
```

As shown in Figure 14, these statements contain two of the participant roles (`substanceOf` and `containerOf`) for participant slots (`?sub` and `?con`, respectively) of `ContainedFluid`, so the system makes the two recursive procedure calls:

***abductive-mf-instantiation***(`ContainedFluid`,  $R = \{\langle ?sub, \text{Blood} \rangle, \langle ?con, \text{heart} \rangle\}$ ,  $S, D$ )

***abductive-mf-instantiation***(`ContainedFluid`,  $R = \{\langle ?sub, \text{Blood} \rangle, \langle ?con, \text{body} \rangle\}$ ,  $S, D$ )

---

**Abductive model formulation**


---

```

procedure instantiate-fragments-with-consequence (proposition  $p$ , domain  $D$ , scenario  $S$ )
  let  $F = \text{query } D \text{ for model fragments with some consequence that unifies with } p$ 
  for each  $f$  in  $F$ :
    for each consequence  $c$  of  $f$  that unifies with  $p$ :
      let  $B = \text{bindings-between}(c, p)$ 
      abductive-mf-instantiation( $f, B, D, S$ )

procedure abductive-mf-instantiation (model fragment  $m$ , role bindings  $R$ , domain  $D$ , scenario  $S$ )
  // Find participant collections  $\{\langle \text{slot}_0, \text{coll}_0 \rangle, \dots, \langle \text{slot}_n, \text{coll}_n \rangle\}$  of  $m$ .
  let  $C_m = \text{participant-collections-of}(m)$ 
  // Find the constraints of  $m$ .
  let  $S_m = \text{constraints-of}(m)$ 
  // Replace variable slots with known entities in constraints & participants
  set  $S_m = \text{replace slot with ent in } S_m \text{ for every } \langle \text{slot}, \text{ent} \rangle \in R$ 
  set  $C_m = \text{replace slot with ent in } C_m \text{ for every } \langle \text{slot}, \text{ent} \rangle \in R$ 
  // If a participant is a model fragment, instantiate it recursively.
  let  $F = \{\langle \text{slot}, \text{coll} \rangle \in C_m : \text{query } D \text{ for } (\text{isa } \text{coll } \text{ModelFragment})\}$ 
  for each  $\langle \text{slot}, \text{coll} \rangle$  in  $F$ :
    // Using the local constraints  $S_m$ , find participant bindings for the recursive call.
    let  $S_f = \text{ground statements in } S_m \text{ which:}$ 
      1. have a participant role of  $\text{coll}$  as its predicate and
      2. have  $\text{slot}$  as a first argument.
    let  $R_f = \text{bindings between participant slots of } \text{coll} \text{ and corresponding entities in } S_f$ 
    // Make a recursive call to instantiate the participant.
    abductive-mf-instantiation( $\text{coll}, R_f, D$ )
  // Find all instance bindings of  $m$  in  $D$ , including ones missing participants
  let  $\text{Instances} = \text{query } D \text{ for bindings of } S_m \wedge C_m$ 
  for each  $I$  in  $\text{Instances}$ :
    // Assume the existence of all unknown participants.
    let  $\text{UnkParticipants} = \{\langle \text{slot}, \text{ent}, \text{coll} \rangle \in I : \text{variable}(\text{ent})\}$ 
    for each  $\langle \text{slot}, \text{ent}, \text{coll} \rangle$  in  $\text{UnkParticipants}$ :
      let  $e = \text{new-skolem-entity}(e, \text{coll})$ 
      set  $I = \text{replace } \langle \text{slot}, \text{ent}, \text{coll} \rangle \text{ with } \langle \text{slot}, e, \text{coll} \rangle \text{ in } I$ 
    // Add the constraints, conditions, consequences, and participant roles to  $\mathbb{B}$ ,
    // and create justifications for this model fragment's instantiation and activation.
    instantiate-model-fragment( $m, I$ )

```

---

**Figure 17: Pseudo-code for abductive model instantiation**

These recursive invocations find no model fragments that can be participants ( $?_{\text{sub}}$  or  $?_{\text{con}}$ ) of `ContainedFluid`. The procedure finds all possible instances of `ContainedFluid` using the bindings  $R$  that obey the constraints (e.g., `(physicallyContains heart Blood)`) and participant types (e.g., `(isa Blood StuffType)`) in  $S$  and instantiates them. Both

recursive invocations instantiate a single `ContainedFluid` instance: one for `heart` and one for `body`. The following assertions are added to  $S$  and to the provisional belief microtheory  $\mathbb{B}$ :

```
(isa mfi0 ContainedFluid)
(substanceOf mfi0 Blood)
(containerOf mfi0 heart)
```

```
(isa mfi1 ContainedFluid)
(substanceOf mfi1 Blood)
(containerOf mfi1 body)
```

These beliefs are plotted as  $f_{6-8}$  and  $f_{9-11}$  in Figure 16. Execution returns to the top-level call to *abductive-mf-instantiation*, where the procedure queries for remaining `FluidFlow` participants. Based on the information in  $S$  – including the model fragment instances that have just been added – the procedure can bind more of the `FluidFlow` participants:  $\{\langle ?source-con, heart \rangle, \langle ?sink-con, body \rangle, \langle ?sub, Blood \rangle, \langle ?source, mfi0 \rangle, \langle ?sink, mfi1 \rangle, \langle ?path, ?path \rangle\}$ . Note that it is still incomplete since there is no entity from the scenario that binds to the `?path` entity. This is because there is no entity in  $S$  is a `Path-Generic` and that satisfies the `FluidFlow` constraint `(permitsFlow ?path Blood heart body)`. In this case, the model fragment is still instantiated. A new symbol (e.g., `mfi2`) is created for the model fragment instance, and unbound entities such as `?path` are assumed and represented with a *skolem term* such as `(SkolemParticipant mfi2 along-Path)`. This skolem term indicates that this entity was assumed as a participant of `mfi2` for the role `along-Path`. The following two assertions are added to  $S$  and to  $\mathbb{B}$ :



```
(isa (SkolemParticipant mfi2 along-Path) Path-Generic)

(permitsFlow (SkolemParticipant mfi2 along-Path) Blood heart body)
```

These beliefs are plotted as  $f_{6-8}$  in Figure 16(b). Notice that this entity is described only in the middle (provisional belief  $\mathbb{B}$ ) layer, since it was generated from model formulation and not from a scenario (i.e., observation, interaction, or instruction). It can be used like any other entity and may be a participant of model fragment instances in subsequent calls.

Once the procedure has a complete list of ground participants, it creates a single instance `mfi2` of type `FluidFlow` and uses this instance to justify the explanandum `naiveH2B`. In other cases, there may be multiple instances that justify the explanandum – consider, for instance, that the agent knew about two `Path-Generic` instances that permit flow of `Blood` from `heart` to `body`. In this case, the agent would not have assumed a `?path` participant but would instead create a `FluidFlow` instance for each path. The instance `mfi2` is described with the following statements:

```
(outOf-UnderSpecifiedContainer mfi2 heart)

(into-UnderSpecifiedContainer mfi2 body)

(substanceOf mfi2 Blood)

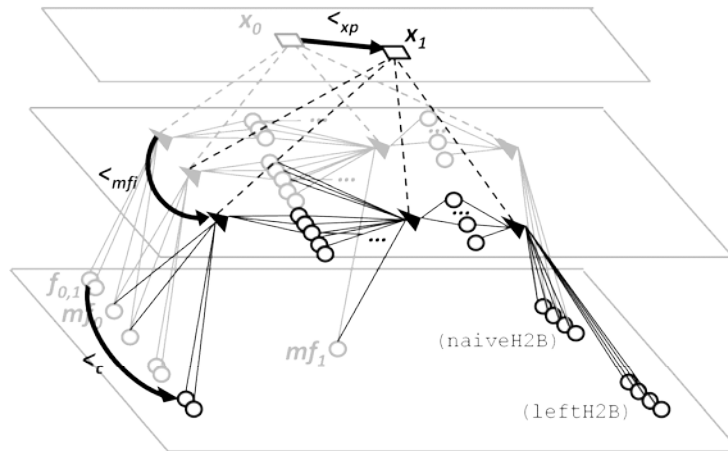
(fromLocation mfi2 mfi0)

(toLocation mfi2 mfi1)

(along-Path mfi2 (SkolemParticipant mfi2 along-Path))
```

All model fragment instantiations and model fragment activations are stored as justifications, and the associated beliefs are stored in  $\mathbb{B}$ . This comprises the entire middle (justification structure) tier in Figure 16(b), which contains in a single well-founded explanation for the explanandum `naiveH2B`. This well-founded explanation has been reified as an explanation node  $x_0$  plotted in the top tier of Figure 16(b). This is the product of the explanation construction algorithm.

Now suppose that the agent learns additional details: (1) the heart is divided into left and right sides (`l-heart` and `r-heart`, respectively) and (2) there is a transfer `leftH2B` of blood from `l-heart` to body. The agent can construct an explanation for `leftH2B` analogous to the process for `naiveH2B`. A new `FluidFlow` instance must be created for `leftH2B`, but the `ContainedFluid` instance for `body` can be reused as its `?sink` participant. After explaining `leftH2B`, the network will resemble Figure 18. There are three important items of note in



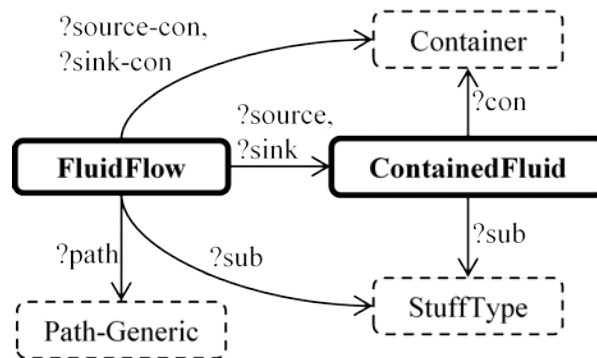
**Figure 18: The network after two explanations have been constructed via abductive model formulation:  $x_0$  explains `naiveH2B`, and  $x_1$  explains `naiveH2B` and `leftH2B`.**

Figure 18, which we will discuss in later sections of this chapter: (1) the new explanation  $x_1$  explains `leftH2B` and also `naiveH2B` (since the `l-heart` is a more specific region of heart), so  $x_0$  and  $x_1$  are in competition; (2)  $x_0$  and  $x_1$  use different but overlapping sets of beliefs and

justifications; and (3) preferences (represented as arrows) have been encoded between concepts  $\leq_c$ , justifications  $\leq_{mf}$ , and explanations  $\leq_{xp}$ . Constructing a new explanation does not eliminate previous explanations; rather, it uses the product of previous explanations to build new structures.

Our abductive model formulation algorithm is exhaustive and complete relative to the scenario  $S$ , domain theory  $D$ , and explanandum  $m$ . It is incomplete with respect to  $S$  and  $D$  alone, since  $m$  guides the recursive search for model fragment instances. For example, the beliefs `(isa lvr Liver)` and `(physicallyContains lvr Blood)` might have been in the scenario  $S$ , but a corresponding `ContainedFluid` would not have been instantiated over  $\{\langle ?sub, Blood \rangle, \langle ?con, lvr \rangle\}$  because the explanandum `naiveH2B` constrained the source and sink containers to `heart` and `body`, respectively.

The abductive model formulation algorithm results in  $(m+e)^p$  model instantiation attempts in the worst case, where  $m$  is the number of models in the domain theory,  $e$  is the number of entities in the scenario, and  $p$  is the number of participant slots per model. The algorithm is guaranteed to converge, assuming that there is no cycle in model fragment dependency. Figure 19 illustrates the dependency graph for the above abductive model formulation example for `naiveH2B`. Each



**Figure 19: A graph of the relationships between model fragments and other collections in the circulatory system example.**

box is a model fragment (bold-bordered) or ordinary collections (dashed) and edges represent participant relationships between types. For instance, the `?source` participant slot of `FluidFlow` requires a `ContainedFluid`, and the `?con` participant slot of a `ContainedFluid` requires a `Container`. Each edge between two model fragments represents a single recursive invocation, so – in the example above – there are two recursive invocations for `FluidFlow`: one for `?source` and one for `?sink`. The algorithm is guaranteed to terminate if it satisfies two constraints:

1. There is no path in the graph from a model fragment  $m$  to a type  $t$  such that  $t$  is equal to  $m$  or is a superordinate of  $m$  in the `genls` hierarchy (see Chapter 2 for the definition of `genls` within an ontological hierarchy).
2. Each model fragment has a finite number of participant slots (i.e., it is graphable with a finite number of nodes).
3. The consequences of the model fragments do not introduce new entities that are not already included as a participant.

To illustrate the necessity of the first constraint, consider what might happen if `Container` was (mistakenly) marked as a `genls` (superordinate) of `FluidFlow` while explaining `naiveH2B`:

1. A call to ***abductive-mf-instantiation*** attempts to model a `FluidFlow`.
2. A recursive invocation of ***abductive-mf-instantiation*** attempts to instantiate a `ContainedFluid` to fill the `?source` participant slot.

3. Since the participant ?con of ContainedFluid can be modeled by a FluidFlow (i.e., (genls FluidFlow Container)), there would be a recursive invocation of *abductive-mf-instantiation* to attempt to instantiate a FluidFlow. Return to (1).

The second constraint is intuitive: if there are infinite participants of a model fragment, there may be infinite recursive invocations to instantiate these participants.

These are reasonable constraints for a domain theory. Aside from guaranteeing convergence, the first constraint guarantees that the resulting scenario model will be *well-founded*, according to Forbus' (1992) formal definition. We include a preprocessing step that ensures that the domain theory  $D$  satisfies this constraint.

The explanations produced by the algorithm contain more detail than everyday verbal explanations (i.e., they decompose phenomena into elementary concepts and causes). In this dissertation, explanations are constructed to promote learning and to answer questions for experimental evaluation, not for inter-agent communication. The problem of constructing explanations for another agent is best addressed elsewhere, since (1) communicating an explanation may have task-specific aspects, and (2) explaining to another person involves knowing what she believes and often including only beliefs and rationale that she lacks.

One problem we have not yet addressed is the problem of *multiple* explanations: after well-founded explanations have been reified as explanation nodes (e.g.,  $x_0$  and  $x_1$  in Figure 18), there frequently exist multiple explanations for a single explanandum (e.g., naiveH2B in Figure 18). Explanation competition and the resolution these competitions are topics of discussion later in this chapter.

#### 4.4.1 Psychological assumptions of explanation construction

Here we discuss the psychological assumptions underlying our abductive model formulation algorithm that were not addressed in Chapter 1.

Our abductive model formulation starts with the explanandum and works backwards to search over a subset of the domain knowledge. A more complete model formulation algorithm would start from all known entities and work forward to instantiate and activate model fragments. Since our model uses a directed backward search, it assumes that people do not consult all of their knowledge when constructing explanations. This is supported by interview transcripts (e.g., in Sherin et al., 2012) where students must be reminded of information they have previously encountered before realizing their explanations are inconsistent. In section 4.5, we discuss how similarity-based retrieval is used to retrieve and reuse previous explanations. This further reduces the space of domain knowledge that is searched during model formulation.

Our algorithm instantiates all possible models that conform to an initial specification and then segments the resulting structure into multiple explanations. This does not seem to be the case for people; the same students in Sherin et al. (2012) appear to construct a single explanation incrementally and only consider an alternative explanation once their initial explanation proves inadequate. In Chapter 9, we discuss opportunities for making our algorithm more incremental and interleaving meta-level analysis.

Our algorithm terminates once an explanandum has been grounded in non-model-fragment types, meaning that termination rests solely on (1) what the agent knows about the scenario/situation, and (2) the model fragments that are available. Consequently, the agent will continue decomposing causes and mechanisms insofar as the scenario permits. This is an unlikely psychological assumption, since it predicts that people will take longer to construct

mechanism-based explanations as they accrue more detailed knowledge about mechanisms. We might remove this assumption by using *modeling assumptions* (Falkenhainer and Forbus, 1991) to limit the types of model fragments considered – and thereby the detail of the qualitative model – based on task- and domain-level properties. Another possible solution is using analogy to infer explanation structure from one case to another. We discuss these ideas further in Chapter 9.

#### 4.4.2 Explanation competition

We have established that conceptual change involves entertaining conflicting ideas. In the two micro-examples of conceptual change in this chapter, we see two different examples of conflict: (1) between two models of the circulatory system and (2) between two different quantities that represent force. We have already described how explanations are constructed. This section describes how explanations compete, and how they are used to organize information.

As shown above, there can be multiple explanations for the same explanandum  $M_i$ . For example, Figure 18 shows the network with two explanations: (1) an explanation  $x_0 = \langle J_0, B_0, M_0 = \{\text{naiveH2B}\} \rangle$  of  $\text{naiveH2B}$  and (2) an explanation  $x_I = \langle J_I, B_I, M_I = \{\text{naiveH2B}, \text{leftH2B}\} \rangle$  of both  $\text{naiveH2B}$  and  $\text{leftH2B}$ . We say that two explanations *compete* over some explanandum(s)  $M$  if and only if they both explain those explanandums. For example,  $x_0$  and  $x_I$  compete to explain  $\text{naiveH2B}$  since  $M_0 \cap M_I = \{\text{naiveH2B}\}$ . By contrast, there is no competition for  $\text{leftH2B}$  since  $x_I$  is its sole explanation.

Explanation competition is important because it indicates a conflict between two different lines of reasoning. In the circulatory system micro-example,  $\text{naiveH2B}$  is explained using knowledge of the heart (via  $x_0$ ) and also knowledge of the left heart (via  $x_I$ ). This is not a serious conflict: one line of reasoning ( $x_I$ ) is just slightly more specific than the other ( $x_0$ ). However,

there can only be one *preferred explanation* per explanandum. The beliefs in preferred explanations – including their assumptions, model fragment instances, and inferences – are *adopted* by the agent, meaning that they are believed. If  $x_I$  becomes the preferred explanation for both `naiveH2B` and `leftH2B` and  $x_0$  is not preferred for any explanandum, then the content of  $x_I$  will be adopted by the agent and the content exclusive to  $x_0$  will not.

We can formalize the mapping from explanandums to their preferred explanation with an *explanation mapping*  $\mathbb{E} = \{\langle m_0, x_0 \rangle, \dots, \langle m_n, x_n \rangle\}$  which maps each explanandum  $m_i$  to its preferred explanation  $x_i$ . The mapping  $\mathbb{E}$  is exhaustive over explanandums, but not exhaustive over explanations (i.e., a single explanation may be preferred for zero or more explanandums). We discuss how explanation preferences are computed later in this chapter.

The explanation mapping plays two important roles in our model of conceptual change. First, it determines, in part, what the agent does and does not believe. For any given belief, if the belief is in some explanation within the explanation mapping, the agent is justified in believing it.

The second role of the explanation mapping is directing the search for knowledge when a new explanandum must be explained. It helps build the  $S$  and  $D$  contexts for the abductive model formulation algorithm discussed above. This means that the content of preferred explanations – and not their non-preferred competitors – is potentially reused in new explanations.

#### 4.5 Explanation retrieval and reuse

Suppose that the agent is asked to explain some explanandum  $m$  (e.g., how blood gets from the heart to the body) on a questionnaire and  $m$  has already been explained by the agent. How would



the agent go about explaining  $m$ ? Since  $m$  has already been explained, the explanation mapping  $\mathbb{E}$  already associates  $m$  with its preferred explanation  $x = \langle J, B, M \rangle$  and the processes and assumptions underlying  $m$  are available in  $x$ 's beliefs  $B$ . This is the simplest case of retrieving and reusing a previous explanation.

But how would the agent explain  $m$  if it had not been previously encountered? Before constructing a new explanation using the abductive model formulation algorithm, the agent must first define the scenario  $S$  and domain theory  $D$  contexts. One simple solution is to define  $D$  as all known model fragments in  $\mathbb{D}$  and define  $S$  as all beliefs in  $\mathbb{D}$  and  $\mathbb{B}$ . This would guarantee that the agent has access to all of the relevant information that it has ever encountered; however, we must also take efficiency into consideration. If we increase the information in  $S$  and  $D$  (e.g., by filling them with all of the agent's knowledge) we will potentially increase the number of recursive calls during model formulation, and we will certainly increase search time. Performance would therefore degrade as the agent accrues knowledge, leading to a *utility problem* (Minton, 1990), which we briefly discussed in Chapter 3. Our solution is to automatically build  $S$  and  $D$  from the contents of previous explanations, which we described above in section 4.4. This does not guarantee that the agent has access to all relevant information, but we do not assume that people have this psychological capability.

Given a new explanandum  $m$  or a new scenario microtheory  $M_s$ , the agent builds model formulation contexts ( $S$  and  $D$ ) from previous explanations, as described above. There are two separate procedures for retrieving previous explanations, shown in Figure 20: (1) ***find-relevant-explanations-for-scenario*** is used when the agent encounters a new scenario microtheory such as a comic graph and (2) ***find-relevant-explanations-for-explanandum*** is used when only an explanandum or query is provided, without an accompanying scenario microtheory.

---

**Similarity-based retrieval of explanations from situations and cases.**

---

**procedure *find-relevant-explanations-for-explanandum* (explanandum  $m$ )**

// Use MAC/FAC to find similar explanandums in  $\mathbb{M}$ , using  $m$  as a probe.

**let**  $SimExplanandums = \text{macfac}(m, \mathbb{M})$

// Return the explanation mappings for the similar explanandums.

**return**  $\{\langle m', x \rangle \in \mathbb{E} : m' \in SimExplanandums\}$

**procedure *find-relevant-explanations-for-scenario* (microtheory  $M_s$ )**

// The case library is all scenario microtheories of previous explanandums

**let**  $CaseLib = All\_scenario\_microtheories - M_s$

// Use MAC/FAC to find similar cases, using  $M_s$  as a probe.

**let**  $SimMicrotheories = \text{macfac}(M_s, CaseLib)$

// Find the explanandums for these similar microtheories.

**let**  $Explanandums = \{m' \in \mathbb{M} : \text{microtheoryOf}(m') \in SimMicrotheories\}$

// Return the explanation mappings for these explanandums.

**return**  $\{\langle m', x \rangle \in \mathbb{E} : m' \in Explanandums\}$

---

**Figure 20: Pseudo-code for best explanation retrieval algorithms, which use MAC/FAC to find explanations that are relevant for a given explanandum or case.**

The procedure ***find-relevant-explanations-for-scenario*** uses MAC/FAC to retrieve previous scenario microtheories that are similar to the new scenario. It then returns all preferred explanations of the explanandums in these similar scenario microtheories. Similarly, the procedure ***find-relevant-explanations-for-explanandum*** retrieves similar explanandums to the new explanandum and then returns all preferred explanations of the similar explanandums.

Once the agent has retrieved a set  $X$  of explanations, it can construct  $D$  as the union of model fragments used in  $X$ , and  $S$  as the union of beliefs in  $M_s$  and all beliefs  $B$  in explanations  $\langle J, B, M \rangle \in X$ . We call this *preferred explanation reuse*. If no previous explanations exist, or if no explanations can be constructed by binding  $S$  and  $D$  in this fashion, then the system sets  $S$  to  $\mathbb{D}$ , and  $D$  to the set of all model fragments in  $\mathbb{D}$ . The simulations in Chapters 6, 7, and 8 use this general pattern for building the  $S$  and  $D$  contexts for model formulation.

Using preferred explanations to seed new explanations has a side effect: the contents of preferred explanations are propagated to new contexts, and the contents of non-preferred explanations are not. This is a positive feedback cycle: if an explanation is preferred, its contents are more likely to be reused, which makes the contents more likely to be part of a *new* preferred explanation.

So far, we have described several characteristics of explanations in our cognitive model: the process by which they are constructed; how they organize information; how they coexist and compete; and how they are retrieved and reused to explain new phenomena. Next we discuss reasoning processes for evaluating explanations and calculating preferences.

#### 4.6 Finding the preferred explanation

The simulations described in this dissertation use the above network structure to organize knowledge and aggregate explanations, but they use two different methods of computing preferred explanations:

1. *Epistemic preferences* are preferential relations over explanations and domain knowledge. They are computed using logical rules and stored as statements in

metaknowledge. The preference ordering over a set of competing explanations is used to determine which is preferred for an explanandum.

2. *Cost functions* map each explanation to a real number indicating its absolute suitability, given what is already believed in the adopted domain knowledge microtheory  $\mathbb{D}_a$  and in other preferred explanations. The best explanation is the one with the lowest numerical cost.

Chapter 6 describes a simulation that uses a cost function, and Chapters 7 and 8 describe simulations that use epistemic preferences. We discuss these in the following sections, and we include ideas for integrating these two approaches in the conclusion of this dissertation.

#### 4.6.1 Rule-based epistemic preferences

Sometimes a model fragment or entity from one explanation can be objectively compared to a model fragment or entity in another explanation, and this helps decide which explanation is better. For instance, the entity *left-heart* – comprised of the *left-atrium* and *left-ventricle* – is objectively more specific than the entity *heart*. If a rule in  $\mathbb{D}$  states that “if  $x$  is a sub-region of  $y$ , then  $x$  is more specific than  $y$ ,” then the agent can encode a specificity-based epistemic preference for *left-heart* over *heart*.

In our model, an epistemic preference (hereafter “preference”) is a binary relation over two units of knowledge. Each preference  $a <_t^d b$  indicates that knowledge  $b$  is strictly preferred to knowledge  $a$  along dimension  $d$  (e.g., specificity, in the above example) over knowledge type  $t$  (i.e., concepts  $c$ , model fragment instances  $mfi$ , or explanations  $xp$ ). The preference between *left-heart* and *heart* entities is shown in Figure 13(b) as a preference between concept-level beliefs

$<_c$ . To be more specific, we would write  $(\text{isa heart Heart}) <_c^s (\text{isa l-heart } (\text{LeftRegionFn Heart}))$ . The dimensions of preference used in our simulations include: specificity ( $s$ ); instructional support ( $i$ ); existence prior to instruction ( $n$ ); completeness ( $c$ ); and revision ( $r$ ). We discuss the criteria for computing preferences in these dimensions below.

Preferences  $b_1 <_c b_2$  between concepts (i.e., beliefs, model fragments, or quantity specifications)  $b_1$  and  $b_2$  are computed via logical criteria. Importantly, if  $b_1$  and  $b_2$  are identical or comparable for specificity (i.e.,  $b_1 \leq_c^s b_2$  or  $b_2 \leq_c^s b_1$ ), we say they are *s-comparable*. The term “commensurable” might apply here as well, but we have already defined it in Kuhn’s (1962) and Carey’s (2009) terms and avoid it here to reduce confusion. Criteria for concept-level preferences are as follows:

<i>Preference</i>	<i>Encoded if and only if</i>
$b_1 <_c^s b_2$	Belief or model fragment $b_1$ is <b>more specific</b> than $b_2$ as inferred by some rule(s) in the domain theory $\mathbb{D}$ .
$b_1 <_c^i b_2$	$b_1$ and $b_2$ are s-comparable; $b_1$ is <b>supported by instruction</b> and $b_2$ is not.
$b_1 <_c^n b_2$	$b_1$ and $b_2$ are s-comparable; $b_1$ is <b>prior knowledge</b> (i.e., believed prior to instruction) and $b_2$ is not.
$b_1 <_c^r b_2$	$b_1$ and $b_2$ are model fragments or quantity specifications, and $b_2$ is a heuristic-based <b>revision</b> of $b_1$ (see section 8.2 for details).

Provided concept-level preferences  $<_c$  over domain knowledge, a preference  $i_1 <_{mfi} i_2$  between model fragment instances  $i_1$  and  $i_2$  is derived from them. These are largely influenced by concept-level preferences  $<_c$ .

<i>Preference</i>	<i>Encoded if all of the following criteria are true</i>
$i_1 <_{mfi}^{d \in \{s, i, n, r\}} i_2$	<ul style="list-style-type: none"> <li><math>i_1</math> and <math>i_2</math> are instances of the same model fragment.</li> <li>At least one <math>i_2</math> participant is preferred <math>&lt;_c^d</math> or <math>&lt;_{mfi}^d</math> to the same-slot <math>i_1</math> participant and all other participants are s-comparable.</li> <li>No <math>i_1</math> participant is strictly preferred <math>&lt;_c^d</math> or <math>&lt;_{mfi}^d</math> to the same-slot <math>i_2</math> participant in the same dimension <math>d</math> as the previous criterion.</li> </ul>
$i_1 <_{mfi}^{d \in \{s, i, n, r\}} i_2$	<ul style="list-style-type: none"> <li><math>i_1</math> and <math>i_2</math> are instances of model fragments <math>m_1</math> and <math>m_2</math>, respectively.</li> <li><math>m_1 &lt;_c^d m_2</math> (i.e., the model fragment of <math>i_2</math> is preferred to that of <math>i_1</math>).</li> <li>All participants of <math>i_2</math> are either identical or preferred <math>&lt;_c^d</math> to the same-slot participants of <math>i_1</math> in the same dimension <math>d</math> as the previous criterion.</li> </ul>
$i_1 <_{mfi}^c i_2$	<ul style="list-style-type: none"> <li><math>i_2</math> is more <b>complete</b> than <math>i_1</math>: <math>i_1</math> contains at least one assumed participant,<sup>23</sup> and one or more of the same-slot <math>i_2</math> participants are not assumed.</li> <li>All non-assumed same-slot participants of <math>i_1</math> and <math>i_2</math> are s-comparable.</li> </ul>

<sup>23</sup> Assumed participants are represented with skolem terms (e.g., (SkolemParticipantFn mfi2 along-Path) ) and not with entities from the scenario (e.g., heart or l-heart). We discussed the conditions for assuming participants in our description of the abductive model formulation algorithm.

Finally, preferences  $<_{xp}$  over explanations are encoded based on preferences  $<_{mfi}$  over model fragment instances.

<i>Preference</i>	<i>Encoded if all of the following criteria are true</i>
$x_1 <_{xp}^{d \in \{s,i,n,r,c\}} x_2$	<ul style="list-style-type: none"> <li>• Explanations <math>x_1</math> and <math>x_2</math> are in competition.</li> <li>• At least one model fragment instance <math>i_2</math> of <math>x_2</math> is preferred to a model fragment instance <math>i_1</math> of <math>x_1</math> such that <math>i_1 &lt;_{mfi}^d i_2</math> and all other model fragments are identical.</li> <li>• No model fragment instance <math>i_1</math> of <math>x_1</math> is preferred to a model fragment instance <math>i_2</math> of <math>x_2</math> such that <math>i_2 &lt;_{mfi}^d i_1</math> over the same dimension <math>d</math> as the previous criterion.</li> </ul>

We have described how preferences over conceptual knowledge (i.e., beliefs, model fragments, and quantity specifications), model fragment instances, and explanations are derived. By these definitions, preferences between concepts  $<_c$  trigger preferences between model fragment instances  $<_{mfi}$ , which in turn trigger preferences  $<_{xp}$  between explanations.

Preferences between explanations decide which explanation is ultimately preferred and mapped in  $\mathbb{E}$ , but this only works if there are no cycles in the explanation preference ordering. Cycles occur when an explanation  $x_0$  is directly or transitively preferred over competing explanation  $x_l$  for one dimension, and  $x_l$  is preferred over  $x_0$  for another dimension. In the mental model transformation example above, consider the agent that starts with knowledge of the heart (i.e., `(isa heart Heart)`) but not the left heart (i.e., `(isa l-heart LeftRegionFn`

Heart))) . Upon learning about the left heart from the equivalent of a textbook, it will have the following specificity, instructional support, and prior knowledge preferences:

$$\begin{aligned} (\text{isa heart Heart}) &<_c^s (\text{isa l-heart (LeftRegionFn Heart)}) \\ (\text{isa heart Heart}) &<_c^i (\text{isa l-heart (LeftRegionFn Heart)}) \\ (\text{isa l-heart (LeftRegionFn Heart)}) &<_c^n (\text{isa heart Heart}) . \end{aligned}$$

If these preferences propagate upward to preferences over model fragment instances and competing explanations, the following preferences over explanations could occur:

$$\begin{aligned} xp_0 &<_{xp}^s xp_1 \\ xp_0 &<_{xp}^i xp_1 \\ xp_1 &<_{xp}^n xp_0 \end{aligned}$$

In Figure 13(b), this cycle in preferences has been reconciled into a single explanation-level preference  $<_{xp}$ . This is achieved with preference aggregation, which we describe next.

### Aggregating epistemic preferences

Epistemic preferences along several dimensions can be aggregated into a single dimension (Doyle, 1991). Our model achieves this with a *preference aggregation function*. The input to the function is a *preference ranking* sequence  $R$  over all dimensions  $D = \{s, i, n, c, r\}$  such as  $R = \langle s, i, n, c, r \rangle$  or  $R = \langle n, c, s, i, r \rangle$ . Informally, the preference ranking describes the relative importance of each dimension of preference, for cycle resolution. The output is a single, acyclic,



partial ordering  $<_{xp}$  over explanations. This is implemented by the following procedure that computes the aggregate ordering  $<_{xp}$ :

```

 $<_{xp} \leftarrow \emptyset$ 

for each  $d \in R$ 

  for each  $pref \in <_{xp}^d$ 

    if  $cycles(<_{xp} + pref) = \emptyset$  then

       $<_{xp} \leftarrow <_{xp} + pref$ 

```

For each dimension of preference, ordered by the preference ranking sequence, all preferences are added to the aggregate ordering unless they result in a cycle in the aggregate ordering. This produces a partial, acyclic ordering over explanations, assuming that preferences  $<_{xp}^d$  in each dimension  $d$  are acyclic. The preference ranking  $R$  thereby influences the decision of which competing explanation is ultimately preferred, which will affect subsequent learning and question-answering.

### Psychological assumptions regarding rule-based epistemic preferences

Here we discuss psychological assumptions underlying our use of rule-based epistemic preferences. Some of the unsupported assumptions of epistemic preferences are resolved by our use of cost functions, which we describe in the next section.

One assumption of our *specificity* preference is that people prefer more specific explanations and concepts over more general ones, all else being equal. This has been common practice in AI for some time (e.g., Poole, 1985). This seems intuitively accurate from an information theoretic

standpoint, since more general information can often be inferred from more specific information (e.g., since the *left-heart* pumps blood to the body, the *heart* pumps blood to the body). Rottman and Keil (2011) show that people attribute more importance to components of an explanation with more elaboration. This specific preference does *not* assume that people prefer to construct more specific explanations when communicating to others, since the explanations we discuss here are self-directed.

Having a *prior knowledge* preference assumes that people may prefer to explain things in terms of entities they are already acquainted with (e.g., the heart) rather than entities that they recently encountered via instruction (e.g., left ventricle). This is indeed the case for students in the control group of Chi et al. (1994a) who (1) explained blood flow in terms of the heart on a pretest, (2) read a textbook passage (twice) which included a description of the left-heart and left-ventricle pumping blood to the body, and (3) still explained blood flow in terms of the heart on the posttest. This is one manner in which we model resistance to change, which is a notable problem in achieving conceptual change (for detailed discussion of resistance, see Feltovich et al., 1994; Chinn and Brewer, 1993).

Our *instructional support* preference assumes that people prefer information that is supported by instruction over comparable information that is not. This is supported by Chi et al. (1994a), who document students changing their mental models when they realize that their beliefs are inconsistent with a textbook passage.

Our *completeness* preference assumes that people prefer explanations that make fewer existence assumptions, all else being equal. We have defined an assumption as a statement that is not readily observed or justified, so all else being equal, assumptions increase uncertainty and decrease the simplicity of an explanation. Lombrozo (2007) provides evidence that people prefer

simpler explanations, and that they believe simpler explanations to be more probable, all else being equal.

Epistemic preferences describe one-dimensional dominance between concepts, model fragments, and explanations. They are sufficient for simulating the conceptual changes described in Chapters 7 and 8, but we do not assume that this is a complete model of psychological explanation evaluation. People have other criteria by which they judge causal explanations, including causal simplicity,<sup>24</sup> coverage of observations, goal appeal, and narrative structure (Lombrozo, 2011). We next discuss how a cost function – used in the simulation in Chapter 6 – can capture some of these macro-level qualities.

#### 4.6.2 Cost functions

In many cases, preferences over individual concepts cannot sufficiently capture what makes one explanation better than another. There are many other considerations when evaluating an explanation: How simple is it? How does it cohere with other explanations I’ve constructed? Does it have consistent causal structure? Our cost function – used in the simulation in Chapter 6 to compute explanation preferences – is designed to answer these questions. In this section we describe the cost function and the elements of explanations that incur costs.

A cost function is a numerical rating of the additional complexity that an explanation would incur the agent. It computes this by summing the cost of *epistemic artifacts* that would be incurred by accepting an explanation (i.e., mapping an explanandum to it in  $\mathbb{E}$ ). Epistemic artifacts (hereafter “artifacts”) include assumptions, contradictions, quantity changes, model

---

<sup>24</sup> Lombrozo (2011) describes simplicity as perceived probability, but it has also been formulated as the minimization of assumptions (Ng & Mooney, 1992) or the minimization of assumption cost (Charniak & Shimony, 1990).

fragments, and more (in table form, below). If an artifact within an explanation, e.g., an assumption, is already used within another preferred explanation in  $\mathbb{E}$ , that artifact does not add to the cost of the explanation in question. When multiple explanations compete to explain an explanandum  $m$ , the minimum-cost explanation  $x$  is chosen as the preferred explanation so that  $\langle m, x \rangle$  is added to  $\mathbb{E}$ . Next we catalog the types of explanation artifacts and describe how explanation costs are computed.

Each artifact is identified by domain-general rules and patterns, and each has a numerical cost. The cost of an explanation  $x = \langle J, B, M \rangle$  is computed as the cost of all new artifacts that would be incurred by accepting  $x$ 's beliefs  $B$ . For instance,  $B$  may contain new assumptions, new model fragment instances, and new beliefs that contradict beliefs in adopted domain knowledge  $\mathbb{D}_a$  or in preferred explanations in  $\mathbb{E}$ . As mentioned above, only *new* artifacts incur a cost, so there is a strong bias for explaining new explanandums with pre-existing assumptions and mechanisms.

Each artifact  $a$  is uniquely defined by the tuple  $a = \langle t_a, B_a \rangle$ , where

- $t_a$  is the *artifact type* (e.g., **Assumption**), which determines the cost of  $a$ . Types and associated costs are listed below.
- $B_a$  is a set of requisite beliefs, such that the cost of  $a$  is incurred if and only if all  $B_a$  are believed (i.e.,  $B_a$  is a subset of the union of  $\mathbb{D}_a$  and the beliefs of all preferred explanations in  $\mathbb{E}$ ).

We use this notation to describe artifacts in Chapter 7.

Let  $\mathbb{A} = \{a_0, \dots, a_n\}$  be the set of all artifacts and let  $\mathbb{A}_i \subseteq \mathbb{A}$  the set of incurred artifacts (i.e., whose costs are incurred by the agent). An artifact is a member of  $\mathbb{A}_i$  exactly when each of its requisite beliefs in  $B_a$  is in  $\mathbb{D}_a$  or in some preferred explanation in  $\mathbb{E}$ . For ease of discussion, we can define the union of adopted beliefs of the agent  $\mathbb{U}_a$  as all beliefs in the adopted domain theory and in preferred explanations:

$$\mathbb{U}_a = \mathbb{D}_a \cup \bigcup_{\langle m, x \rangle \in \mathbb{E}} B: x = \langle J, B, M \rangle$$

We can now compute the set of incurred artifacts  $\mathbb{A}_i$  as all artifacts in  $\mathbb{A}$  whose beliefs  $B_a$  are in  $\mathbb{U}_a$ :

$$\mathbb{A}_i = \{\langle t_a, B_a \rangle \in \mathbb{A}: B_a \subseteq \mathbb{U}_a\}$$

We list the artifact types  $t_a$  used in the simulation in Chapter 6, and we describe how requisite beliefs  $B_a$  of each type are computed. Importantly, one type of artifact has a negative cost, so it provides a utility to the agent rather than a penalty.

<b><math>t_a</math>: cost</b>	<b><math>B_a</math> constituents</b>
Contradiction: 100	$B_a$ is any set of beliefs such that the conjunction of beliefs $B_a$ – and no strict subset thereof – is inconsistent.
Asymmetric quantity change: 40	$B_a = \{b\}$ , where $b$ is a statement in an explanation $x$ 's

	metaknowledge $B_m$ that describes a quantity change in $x$ that does not have a reciprocal quantity change in a cyclical state-space. <sup>25</sup>
Assumed quantity change: 30	$B_a = \{b\}$ , where $b$ is an assumed quantity change. These are costly because quantity changes must be explained at some point by introducing a process instance, since processes are the sole mechanism of change in a physical system (Forbus, 1984).
Model fragment: 4	$B_a = \{(isa\ mf\ ModelFragment)\}$ , where $mf$ is a model fragment, e.g., <code>ContainedFluid</code> in the circulatory system micro-example.
Assumption: 3	$B_a = \{b\}$ , where $b$ is an assumed proposition.
Model fragment instance: 2	$B_a = \{(isa\ inst\ mf)\}$ where $inst$ is the instance name and $mf$ is the model fragment type, e.g., <code>(isa mfi0 ContainedFluid)</code> in the circulatory system micro-example.
Credibility: [-1000, 0)	$B_a = \{b\}$ , where $b$ was communicated from another source. The utility (i.e., negative cost) of accepting $b$ is

<sup>25</sup> Asymmetric quantity changes are possible in any cyclic state space, such as the water cycle, the carbon cycle, breathing, the seasons, and day/night. The day/night explanation “night turns to day in Chicago because the earth rotates so that Chicago faces the sun, and day turns to night in Chicago because clouds cover the sun” is asymmetric: there is no mention of how the earth rotates to block Chicago from the sun for the next sunrise. We provide more examples of asymmetric quantity changes in Chapter 6.

---

proportional to the credibility of the source.

---

The artifacts and costs listed above are sufficient for simulating the mental model transformation in Chapter 6, but we do not believe this list is complete. Also, the costs listed above were determined empirically to maximize the accuracy of the simulation in Chapter 6, so we had to make several psychological assumptions which we discuss below.

According to Occam's razor, a simpler explanation is better, all else being equal. The penalties for model fragments and their instances promotes qualitative parsimony (i.e., minimizing the new *kinds* of entities postulated) and quantitative parsimony (i.e., minimizing the

---

### Explanation and belief cost computation

---

```
// Compute the cost that would be incurred by adopting an explanation.
procedure explanation-cost (explanation  $x = \langle J, B, M \rangle$ )
  // Find artifacts  $A_x$  pertaining to  $x$  that are not presently incurred.
  let  $A_x = \{\langle t_a, B_a \rangle \in \mathbb{A}/\mathbb{A}_i : B_a \cap B \neq \emptyset\}$ 
  // Find artifacts  $A$  incurred if  $x$  were adopted. Recall  $\mathbb{U}_a$  is adopted beliefs.
  let  $A = \{\langle t_a, B_a \rangle \in A_x : B_a \in (B \cup \mathbb{U}_a)\}$ 
  // Return the sum of the costs of these artifacts.
  return  $\sum_{a \in A} \text{cost}(a)$ 

// Compute the cost that can be saved by retracting a belief.
procedure retraction-savings (belief  $b$ )
  // If  $b$  is not in a preferred explanation...
  if  $\nexists \langle m, \langle J, B, M \rangle \rangle \in (b \in B)$  then
    // Find artifacts  $A$  supported by  $b$  that are presently incurred.
    let  $A = \{\langle t_a, B_a \rangle \in \mathbb{A}_i : b \in B_a\}$ 
    // Return the sum of the cost of these artifacts.
    return  $-\sum_{a \in A} \text{cost}(a)$ 
  else return 0
```

---

**Figure 21: Pseudo-code for computing an explanation's cost and a belief's cost using a cost function. Note that the cost of any explanation that is presently adopted (i.e., an explanandum is mapped to it in  $\mathbb{E}$ ) is zero.**

*number* of new entities postulated), respectively. Promoting parsimony and penalizing assumptions makes a simpler explanation less costly, all else being equal.

The cost function is used for two purposes in our computational model: (1) computing a preferred explanation from multiple competing explanations and (2) retrospectively changing  $\mathbb{D}_a$  and  $\mathbb{E}$ . In the case of explanation competition, a new explanandum  $m$  (e.g., the changing of Chicago's seasons, in Chapter 6) is explained by the agent, and multiple explanations  $X$  compete to explain  $m$ . The cost function is used to decide which explanation  $x \in X$  to associate with  $m$  as its preferred explanation in  $\mathbb{E}$ . Computing the cost of an explanation  $x$  is equivalent to computing the total cost of all artifacts in  $\mathbb{A}/\mathbb{A}_i$  that would be added to  $\mathbb{A}_i$  if  $\langle m, x \rangle \in \mathbb{E}$ . This algorithm is shown in Figure 21. The agent uses the function *explanation-cost* to find the minimal cost explanation in  $X$ .

The cost function is also used to retrospectively change  $\mathbb{D}_a$  and  $\mathbb{E}$  to reduce cost. For instance, it could be the case that the cost of  $\mathbb{A}_i$  could be significantly reduced by switching the preferred explanation for some explanandum(s) in  $\mathbb{E}$  or removing some belief(s) from  $\mathbb{D}_a$ . Consider the sequence of events in Figure 22 that occurs in a simulation trial in Chapter 6: the agent makes the locally optimal choices for two explanations, but then learns some new information that renders the two explanations mutually inconsistent, despite retaining their individual internal consistency. In this situation, the contradictions may be removed by removing the credible beliefs  $b_0$  or  $b_1$  from  $\mathbb{D}_a$  (thus losing the credibility bonus) or changing the preferred explanation for Chicago's seasons (explanandum  $m_0$ ), Australia's seasons (explanandum  $m_1$ ), or both. Making these changes may alter the set of beliefs  $\mathbb{U}_a$  that the agent holds to be true, so this is a mechanism of belief revision.



What the agent does	In our model:
1. Explained Chicago's difference in summer and winter temperatures (explanandum $m_0$ ) with an explanation $x_0$ of the earth being closest to the sun in Chicago's summer and being furthest from the sun in Chicago's winter.	$m_0 \in \mathbb{M}$ $\langle m_0, x_0 \rangle \in \mathbb{E}$
2. Explained Australia's difference in summer and winter temperatures (explanandum $m_1$ ) with the similar explanation $x_1$ to $x_0$ , using the same mechanisms and assumptions.	$m_1 \in \mathbb{M}$ $\langle m_1, x_1 \rangle \in \mathbb{E}$
3. Learned from a credible source that Australia's winter coincides with Chicago's summer (belief $b_0$ ) and Australia's summer coincides with Chicago's winter (belief $b_1$ ).	$\{b_0, b_1\} \subseteq \mathbb{D}_a$ $\langle \text{Credibility}, \{b_0\} \rangle \in \mathbb{A}_i$ $\langle \text{Credibility}, \{b_1\} \rangle \in \mathbb{A}_i$
4. Detected four contradictions due to $b_0, b_1$ , and the beliefs in $x_0$ and $x_1$ , e.g., the earth cannot be closest to the sun in Chicago's summer and farthest in Australia's winter at the same time, since they temporally coincide.	Four contradiction artifacts added to $\mathbb{A}_i$ , of the form $\langle \text{Contra}, \{b_0, b_1, b_{x_0}, b_{x_1}\} \rangle$ where $b_{x_0}$ and $b_{x_1}$ are beliefs from $x_0$ and $x_1$ , respectively.

**Figure 22: A sequence of events from the simulation in Chapter 6 that produces several contradictions between best explanations and credible domain knowledge.**

Restructuring the entire contents of  $\mathbb{D}_a$  and remapping all explanandums in  $\mathbb{E}$  to find the minimal cost configuration is very costly. This is due to the number of possible mappings in  $\mathbb{E}$  and beliefs in  $\mathbb{D}_a$  that must be considered. If there are  $m$  explanandums with  $x$  explanations each and  $b$  domain beliefs which can be either adopted (i.e., in  $\mathbb{D}_a$ ) or not (i.e., in  $\mathbb{D} \setminus \mathbb{D}_a$ ), then there are  $2^b x^m$  possible configurations. If there are 16 explanations for both Chicago and Australia, this is equivalent to  $2^2 * 16^2 = 1024$  configurations for just the two explanations and two domain beliefs in Figure 22. While this is not a serious problem for this small example, the time complexity is exponential on the number of explanandums being considered. We avoid this combinatorial explosion by using the greedy, local reconstruction algorithm shown in Figure 23.

Our local reconstruction algorithm takes a single artifact as input, finds the domain beliefs

---

### Locally restructuring the KB

---

```

function restructure-around-artifact (artifact  $a = \langle t_a, B_a \rangle$ )
  // Find supporting explanandums
  let  $M_a = \{M_i \in \mathbb{M} : \langle M_i, \langle J, B, M \rangle \rangle \in \mathbb{E} \wedge (B_a \cap B) \neq \emptyset\}$ 
  // Find supporting beliefs in the domain theory
  let  $D_a = \mathbb{D}_a \cap B_a$ 
  // Iterate until no further local revisions are made.
  let revised = true
  while revised:
    set revised = false
    for each  $M_i$  in  $M_a$ :
      // Find explanations that can explain this.
      let  $X = \{\langle J, B, M \rangle \in \mathbb{X} : M_i \in M\}$ 
      // Find the least cost explanation.
      let  $x = \min_{x \in X} \textit{explanation-cost}(x)$ 
      // Make the least cost explanation the best explanation, if not already.
      if  $\langle M_i, x \rangle \notin \mathbb{E}$  then:
        replace  $\langle M_i, * \rangle$  with  $\langle M_i, x \rangle$  in  $\mathbb{E}$ 
        set revised = true
    for each  $d$  in  $D_a$ :
      // If this belief can be retracted to reduce cost, retract it.
      if retraction-savings( $d$ ) > 0 then
        // Remove  $d$  from adopted beliefs.
        set  $\mathbb{D}_a = \mathbb{D}_a - d$ 

```

---

**Figure 23: Algorithm for restructuring knowledge based on the presence of a high-cost artifact.**

and explanations that support it, and greedily reconfigures the beliefs and affected explanandums to reduce the cost. This involves remapping individual explanandums in  $\mathbb{E}$  and adding or removing beliefs in  $\mathbb{D}_a$ . Each explanandum under consideration changes its mapping in  $\mathbb{E}$  to a lower-cost explanation (if there is one), and each belief in  $\mathbb{D}_a$  under consideration is added or removed from  $\mathbb{D}_a$  if it will reduce cost. This occurs in a closed loop, until no single action can reduce the cumulative cost, and then the algorithm terminates. It is guaranteed to terminate,

since each action – and therefore each loop – must reduce the cost of incurred artifacts, and cost can only be finitely reduced.

The series of unilateral changes in the restructuring algorithm is not guaranteed to find the minimum cost configuration; however, the average case performance is much more tractable with respect to the number of beliefs and explanandums considered. Using the same analysis as above, the number of cost computations on each loop is  $2b + xm$ , which equals 36. The number of loops varies with the content of the explanations, and a carefully-engineered scenario could still produce a worst-case performance of  $2^b x^m$  cost computations in total, identical to finding the optimal cost above. In the Figure 22 example from Chapter 6, the algorithm takes a total of two loops to reach a stable configuration. This required 72 cost computations instead of 1024 in the worst case for the same circumstance. The algorithm is not guaranteed to remove the artifact that was provided as input; rather, the input artifact is used as a marker for possible cost optimization.

### **Psychological assumptions regarding cost functions**

Cost functions capture psychological explanation preferences that are not possible using rule-based epistemic preferences alone; however, they make some additional assumptions about how people evaluate explanations.

Several psychological assumptions underlie the types of artifacts that incur a cost in our model. In our model, process instances and quantity changes are the mechanisms and effects of a dynamic system, respectively. These comprise the root and intermediary causes within a system. People prefer explanations with fewer causes, all else being equal (Lombrozo, 2007), so it is sensible to penalize process instances and quantity changes.

By penalizing contradictions, we assume that people desire consistency within and across explanations. This assumption is common to the other theories of conceptual change in Chapter 2, and it is clearly supported in interviews (e.g., Sherin et al., 2012) where students revise their explanations when they detect inconsistencies.

We assume that an explanation's quality is not solely determined by its probability. When we refer to an explanation's probability, we mean the joint probability of the explanation's assumptions relative to other adopted beliefs. To illustrate, here is how we might compute the most preferable explanation using probability alone: we use probabilities to represent the agent's purported likelihood of a given belief, and then search for a maximum a-posteriori (MAP) truth value assignment to all existing assumptions. The explanation that conforms to this set of assumptions would be the preferred explanation. We could then model people's simplicity preference (i.e., minimizing the number of causes, similar to above) by assigning more complex causes a lower prior probability. Finally, we can avoid contradictions by encoding a zero for the joint probability of mutually inconsistent beliefs (e.g.,  $\{b_0, b_1, b_{x0}, b_{x1}\}$  in Figure 22). Thus, when new knowledge causes an explanation, the agent could revise its explanation by searching for more probable truth value assignments for assumptions.

The alternative, purely probabilistic model we have just described makes a very strong assumption that we do not make in our computational model: assignments of truth values to assumptions that are equally probable are equally preferable to people. To illustrate why this is problematic, consider a student with two contradictions ( $\perp_a$  and  $\perp_b$ ) in his adopted beliefs. Since  $\perp_a$  or  $\perp_b$  alone will result in a probability of zero, resolving  $\perp_a$  while  $\perp_b$  still exists does not measurably improve the student's interpretation of the world, so no action need be taken. This is

not to suggest that a purely probabilistic approach to evaluating explanations is infeasible, but it would require additional considerations for evaluating explanations both locally and globally.

Our cost function assigns all assumptions an identical cost, but it is not likely that people view all assumptions as equally desirable. This could be improved by representing the uncertainty of beliefs – potentially using probabilities – and then computing the cost of an assumption as a function of uncertainty. We discuss this further in Chapter 9.

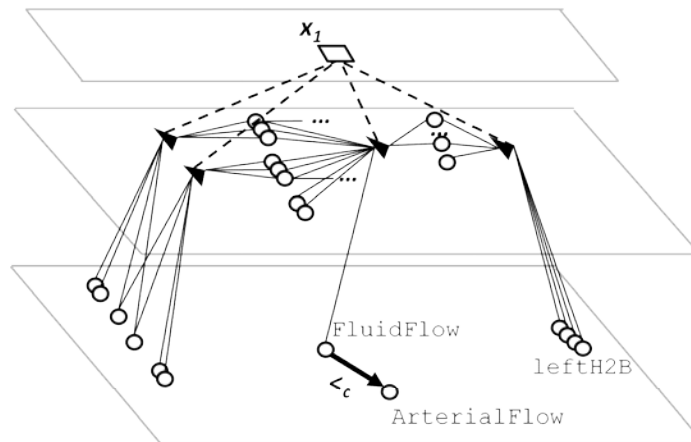
#### 4.7 Retrospective explanation

In our discussion of cost functions, we described a restructuring algorithm that manipulates previously explained beliefs and transitions support to lower-cost explanations. This requires that previous explanations are already present for evaluation and potential transition.

Importantly, beliefs and model fragments may have been added to  $\mathbb{D}$  since an explanandum was encountered, so the agent might be able to construct a better explanation than presently exists.

*Retrospective explanation* is the process of constructing new explanations for previous explanandums.

The first task in retrospective explanation is to detect opportunities for retrospective explanation. Adding knowledge to the domain theory  $\mathbb{D}$  can change the space of possible explanations for an explanandum, but not all expansions of  $\mathbb{D}$  affect all explanandums  $\mathbb{M}$ . In the simulations described in Chapters 7 and 8, concept preferences  $<_c$  dictate opportunities for retrospective explanation.



**Figure 24: Model fragment `ArterialFlow` is preferred over `FluidFlow` due to greater specificity, but `leftH2B` has not yet been explained using the preferred knowledge.**

As illustrated in Figure 24, explanandum `leftH2B` is explained with a model fragment `FluidFlow`, but not with the preferred model fragment `ArterialFlow`. This might occur if `ArterialFlow` is a more specific  $<_c^s$  model fragment, but it was learned after `leftH2B` was explained. A similar pattern could occur when revising models of force and motion: the observation that a ball is rolling to the left has been explained with a model  $m_0$  of force-driven movement, but  $m_0$  has since been revised as  $m_1$  such that  $m_0 <_c^r m_1$ . In both of cases, a preferred model fragment was not present when an explanandum was explained. A retrospective explanation opportunity exists in both of these cases. More generally, a retrospective explanation opportunity exists any time a concept  $c$  has been used to explain an explanandum  $m$  and a preferred concept  $c'$  (i.e.,  $c <_c c'$ ) has not been attempted for use with that explanandum. Every retrospective explanation opportunity will thus be a triple of a belief plus a pair of concepts.

The simulations in Chapters 7-8 search for retrospective explanation opportunities any time concept-level preferences are computed after incorporating a scenario.<sup>26</sup> Once a retrospective explanation opportunity is found, the explanandum is explained using the abductive model formulation algorithm described above. This provides additional support for previously-explained beliefs without disrupting existing explanations in the network. The evaluation techniques described above (i.e., preference computation and cost function) can then be used to determine whether the new explanation is preferable to existing ones. This is how retrospective explanation results in belief revision.

With respect to Figure 24, retrospective explanation may fail to construct a new explanation for `leftH2B` using `ArterialFlow`. In this case, the triple  $\langle \text{leftH2B}, \text{FluidFlow}, \text{ArterialFlow} \rangle$  is stored as a retrospective explanation failure so that the system will not attempt retrospective explanation for the same reason. The existing explanation  $x_7$  will remain the best explanation for `leftH2B`.

The agent may add new information (i.e., beliefs, models, and quantities) via inductive learning (e.g., Chapter 5), instruction (e.g., Chapter 7), or heuristic-based revision (e.g., Chapter 8), but these additions do not by themselves constitute successful conceptual change. The agent ultimately achieves conceptual changes using the methods described in this chapter. After acquiring or revising information, the agent propagates it to new contexts and scenarios by using it to explain new and previous phenomena. If the new explanations are preferable to prior ones, the agent re-justifies its beliefs with new explanations. The agent can thereby adopt new combinations of information and new representations in the presence of conflicting knowledge, which is, by definition, conceptual change.

---

<sup>26</sup> In situations where the agent does not have time to reflect on previous scenarios, retrospective explanation can be delayed until a later time. We discuss the implications of delaying retrospective explanation – and ways to experimentally measure the effects – in section 9.4.

## Chapter 5: Learning intuitive mental models of motion from observation

Conceptual change does not begin with a blank slate. This chapter presents a simulation of how intuitive models can be learned from a sequence of observations. This provides an account of how flawed mental models are formed as a precursor to conceptual change, but it does not in itself constitute conceptual change. Other systems have learned humanlike misconceptions from examples (e.g., Esposito et al., 2000), but with different methods and knowledge representations, as we discuss in Chapter 9.

Students' pre-instructional knowledge has been explored in the cognitive science literature in many domains. This knowledge is also referred to as *preconceptions*, *intuitive theories*, and – when inconsistent with scientific theories – *misconceptions* or *alternate conceptions*. Pre-instructional knowledge in scientific domains (e.g., dynamics and biology) is presumably learned via observation and interaction with the world. The simulation described in this chapter provides a computational account of how descriptive mental models of dynamics might be learned via observations.<sup>27</sup>

We use the term *descriptive mental models* here because the models learned by this simulation describe what-follows-what without specifying conceptual mechanisms and physical processes that cause change. Consider the following system of beliefs:

When an object  $a$  is moving in the direction  $d$  of another object  $b$ :  
 $a$  might touch object  $b$  and push it in direction  $d$ , in which case:  
 $b$  may block  $a$ , or  
 $b$  may move in direction  $d$ .

---

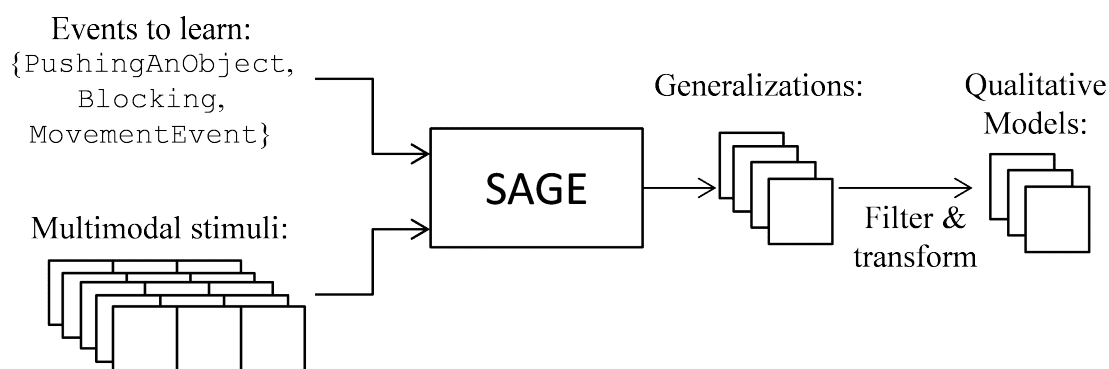
<sup>27</sup> This chapter expands the original account described in Friedman, Taylor, and Forbus (2009).



This simple descriptive account of dynamics is incomplete and it does not appeal to any conceptual quantities such as force, inertia, or impetus,<sup>28</sup> but it still has considerable predictive and explanatory power. It contains temporal constraints (i.e., one state or event follows another in time), it is abstract (i.e., it does not mention specific types of objects such as “ball” or “block”), and it is parameterized (i.e., it can occur for multiple directions  $d$ ).

The structure of this simulation is shown in Figure 25. The input to the system is (1) a set of event types to model and (2) a sequence of scenarios, implemented via microtheories, for learning about this set of events. The system first finds instances of the event types within the stimuli and constructs temporally-encoded cases for each event instance. Next, SAGE is used to construct generalizations of each type of event. These generalizations are subsequently filtered and converted into qualitative models.

To evaluate what is learned, the resulting qualitative models are used on two problem-solving tasks from the learning science literature: one from Brown (1994), and one from the *Force Concept Inventory* (Hestenes et al., 1992). This helps us determine whether the learned qualitative models can simulate the pre-instructional mental models of students. We are



**Figure 25: Topology of the Chapter 5 simulation.**

<sup>28</sup> See Chapter 8 for a simulation that generates and uses conceptual quantities.

principally interested in simulating students' misconceptions – recall that the objective of this simulation is to model how students learn from observation, and students do not arrive at correct Newtonian models using observation alone. This simulation provides evidence to support the first two claims of this dissertation:

***Claim 1:*** Compositional qualitative models provide a consistent computational account of human mental models.

***Claim 2:*** Analogical generalization, as modeled by SAGE, is capable of inducing qualitative models that satisfy Claim 1.

The other simulations provide additional support for Claim 1, but no other simulation provides support for Claim 2. Importantly, the qualitative models learned in this simulation do not describe continuous causal mechanisms. This is because SAGE does not hypothesize causal mechanisms such as processes and quantities where none are already believed to exist. We next describe our simulation, including the training and testing data, the learning processes, and a comparison to human mental models.

### **5.1 Using multimodal training data**

The training data for this simulation is *multimodal* because each training case is created using two different modes of input: sketches and simplified English. This is a simplified

approximation<sup>29</sup> of what a person might encounter in daily experience. Each case contains relational knowledge that CogSketch<sup>30</sup> encoded from hand-sketched comic graphs such as Figure 26. Each case also contains knowledge that the natural language understanding system EA NLU (Tomai & Forbus, 2009) semi-automatically encodes from one or more English sentences that describe the comic graph. The following English sentences accompany the comic graph in Figure 26:

The child *child-15* is here.

The child *child-15* is playing with the truck *truck-15*.

The car *car-15* is here.



**Figure 26: A comic graph stimulus created using CogSketch**

Since cross-modal reference resolution is a difficult open problem, we factor it out by using the internal tokens from the sketch (in italics) within the sentence. One can think of this as providing the same kind of information that a teacher would be giving a child by pointing at objects while talking about them. EA NLU uses the term *child-15* to refer to the *Child* entity that is playing with the *Truck* entity *truck-15*. These are the same entity names used by

<sup>29</sup> See Chapter 3 for a discussion of the psychological assumptions and limitations of using sketches as perceptual output.

<sup>30</sup> See Chapter 3 for a functional overview of CogSketch.

CogSketch. The outputs of CogSketch and EA NLU are automatically combined into a single, coherent scenario microtheory.

For each multimodal scenario microtheory, the simulation finds instances of target concepts (see Figure 25) such as the two `PushingAnObject` instances (blue arrows in the middle frame) and the two `MovementEvent` instances (green arrows in the rightmost frame) in Figure 26. For each instance of each event, e.g., the truck moving in the rightmost frame of Figure 26, the system creates a new microtheory that describes that event. The temporal extent of the event (e.g., the truck moving) is recorded as the `currentState`, and other statements in the comic graph are recorded in the event microtheory, relative to the current state. For example, the event microtheory that describes the truck's movement would contain the following statements:

```
(cotemporal currentState (isa move-truck-15 MovementEvent))
(cotemporal currentState (objectMoving move-truck-15 truck-15))
(cotemporal currentState (motionPathway move-truck-15 Right))
(startsAfterEndingOf currentState (touching truck-15 child-15))
(startsAfterEndingOf currentState (isa push-15-0 PushingAnObject))
(startsAfterEndingOf currentState (providerOfForce push-15-0 child-15))
(startsAfterEndingOf currentState (objectActedOn push-15-0 truck-15))
(startsAfterEndingOf currentState (dir-Pointing push-15-0 Right))
```

The system encodes the truck's rightward movement within in the `currentState`, and this happened right after (i.e., `startsAfterEndingOf`) the child touched the truck and pushed it to the right. These statements alone provide a concise account of cause (e.g., `PushingAnObject`) and effect (i.e., `MovementEvent`); however, these are not the only statements in the event

microtheory. There are many other statements that are irrelevant – or worse, confusing – for learning about cause and effect. These include:

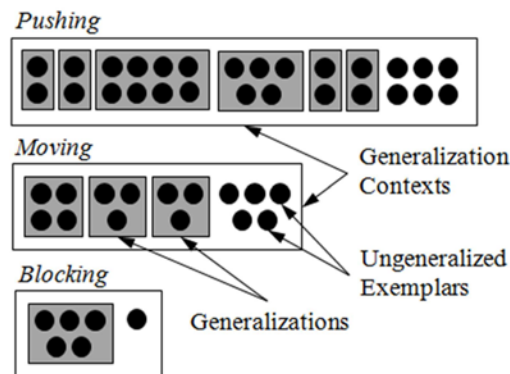
```
(temporallySubsumes (touching truck-15 ground-15) currentState)
(temporallySubsumes (touching car-15 truck-15) currentState)
(temporallySubsumes (touching car-15 ground-15) currentState)
(cotemporal currentState (isa move-car-15 MovementEvent))
(cotemporal currentState (objectMoving move-car-15 car-15))
(cotemporal currentState (motionPathway move-car-15 Right))
```

These irrelevant statements describe the truck touching the car, the car touching the ground, and the car moving simultaneously. There are many more such irrelevant statements that are not shown here, including positional relations, relative sizes and shapes of the glyphs, and more. One important task in learning from observation is distinguishing causally-relevant information from incidental or distracting information. This is done automatically with SAGE (see section 3.4.3), and we address this challenge next.

So far, we have shown how the system finds instances of its target concepts within multimodal scenario microtheories and creates a temporally-encoded microtheory for each instance. The temporal relations help record what might be a cause and what might be an effect of each event, e.g., if movement starts after pushing, then movement is not a plausible cause of pushing, but it is a plausible effect. Temporal relations also add significant relational structure to the representation of the event, which will aid in analogical learning with SAGE. The next section describes how SAGE abstracts the central causal structure of these scenarios from the irrelevant, confusing statements.

## 5.2 Creating generalizations of Pushing, Moving, and Blocking with SAGE

The system maintains a separate SAGE generalization context for each of the event types it is given to learn (see Figure 25). This simulation creates three generalization contexts: one for `PushingAnObject`, one for `Blocking`, and one for `MovementEvent`. Instances of events are added to the generalization context for that event type. For example, each temporally-encoded microtheory that describes a `MovementEvent` instance is added to the `MovementEvent` generalization context – and no other – to be automatically generalized using SAGE.<sup>31</sup>



**Figure 27: The three SAGE generalization contexts after using SAGE to generalize temporally-encoded microtheories about pushing, moving, and blocking.**

The contents of these generalization contexts during a simulation are illustrated in Figure 27. Using a separate SAGE generalization context for each concept prevents SAGE from conflating different concepts during supervised learning. Within each context, however, SAGE may have multiple generalizations. For instance, within the `PushingAnObject` context, there may be a pushing generalization where a `MovementEvent` follows the push, and another pushing generalization where a `Blocking` occurs simultaneously with the push and no `MovementEvent` ensues. This clustering is unsupervised, arising from the properties of the data itself.

<sup>31</sup> The SAGE generalization algorithm is described in Chapter 3.

As discussed in Chapter 3, each SAGE generalization contains a set of statements, and each statement has a probability. Recall that the microtheories given to SAGE in this simulation contain temporal relations between a `currentState` and other events and statements. Consequently, the generalizations produced by SAGE will be probabilistic accounts of what happened before, during, and after the `currentState`. The statements with high probability are more characteristic of the event than low probability statements.

The probabilistic generalizations produced by SAGE are not themselves causal models. However, they contain sufficient temporal and statistical information to create descriptive qualitative models.

### 5.3 Converting SAGE generalizations to qualitative models

This work is the first to construct qualitative models from probabilistic generalizations. SAGE generalizations are converted to qualitative models in two steps: (1) probability filtering and (2) causal assignment. *Probability filtering* involves discarding expressions within the generalization that are below a given probability threshold. This retains expressions that are more probable in the generalization (e.g., that two objects are touching during a push event) and discards expressions that are less probable (e.g., that one of the objects is a toy truck).

<b><i>s</i> relation to event <i>e</i></b>	<b>Roles in model</b>
<i>s</i> starts before <i>e</i> starts	cause
<i>s</i> starts after <i>e</i> starts	effect
<i>s</i> subsumes & starts before <i>e</i>	constraint, cause
<i>s</i> subsumes & starts with <i>e</i>	constraint, cause, effect
<i>s</i> and <i>e</i> are cotemporal	constraint, cause, effect

**Figure 28: Given a statement *s* and its temporal relationship to an event *e*, how to calculate the causal role(s) of *s* in a qualitative model of *e*.**

After low-probability statements are filtered, *causal assignment* determines each remaining statement's causal role with respect to the central event. This is a simple lookup, using the temporal relation(s) between the statement and the `currentState` where the event occurs. The lookup table is shown in Figure 28. Sometimes there is equal evidence that a statement can play multiple roles, such as  $\{\textit{constraint}, \textit{cause}\}$  or  $\{\textit{constraint}, \textit{effect}\}$  or  $\{\textit{constraint}, \textit{cause}, \textit{effect}\}$ . In these cases, the system always chooses *constraint*. To illustrate why this is the case, suppose that our generalization describes object *a* starting to touch object *b* whenever *a* starts to push *b*, but never before and never after. It could be that:

1. touching *causes* pushing,
2. touching is an *effect* of pushing, or
3. touching is a necessary *constraint* for pushing to occur.

The bias for adding touching as a constraint seems intuitive, but it has important implications for the resulting qualitative model. Recall from our discussion of qualitative model fragments in Chapters 3 and 4 that constraints limit the logical applicability of the model fragment. Adding touching as a constraint for pushing – rather than a consequence of pushing – will limit the logical applicability of the model, all else being equal. This means that the model will apply in fewer situations, so some events may go unpredicted or unexplained. However,



limiting the logical applicability of a model also reduces false positives – that is, it will less frequently make erroneous predictions or misattributions of causality.

Model Push05	
Participants:	
?P1 Entity, ?P2 Entity,	
?P3 PushingAnObject,	
?D1 Direction, ?D2 Direction	
Constraints:	
(providerOfForce ?P3 ?P1)	Object $p_1$ touches and pushes object $p_2$ in direction $d_1$ . The direction between $p_1$ and $p_2$ is $d_1$ .
(objectActedOn ?P3 ?P2)	
(dir-Pointing ?P3 ?D1)	
(touching ?P1 ?P2)	
(dirBetween ?P1 ?P2 ?D1)	
(dirBetween ?P2 ?P1 ?D2)	
Consequences:	
(causes	This causes object $p_2$ to travel in the direction $d_1$ of the push.
(active ?self)	
(exists ?M1	
(and (isa ?M1 MovementEvent)	
(objectMoving ?M1 ?P2)	
(motionPathway ?M1 ?D1)))	

**Figure 29: One of the qualitative models learned by the simulation that causally relates pushing and movement. Summaries of constraints and consequences shown at right.**

After causal assignment occurs, every high-probability statement in the SAGE generalization has been assigned a role in a qualitative model. The entities in the constraints are converted to variables and become the participants of the resulting model. This produces *encapsulated histories* (Forbus, 1984), which are descriptive qualitative models that causally or temporally relate events over time. Figure 29 shows one such qualitative model learned by the simulation. It describes a `PushingAnObject` event and several spatial and relational constraints over the objects involved, and a `MovementEvent` occurs as a result. The set of constraint statements are directly imported as constraints of the model, but causes and effects are listed in the consequences of the model. For instance, in Figure 29, the `MovementEvent` ?m1 is an effect of the `PushingAnObject` ?p1, so the following statement is a consequence:

```
(causes
  (active ?self)
  (exists ?M1
    (and (isa ?M1 MovementEvent)
          (objectMoving ?M1 ?P2)
          (motionPathway ?M1 ?D1))))
```

Recall that the constraints and participants of a qualitative model are logical antecedents to the construction and activation of an instance of the model (e.g., (active ?self)) over those participants. For example, when an instance `Push05-Instance` of model `Push05` is created and activated with `?p2` bound to `pushed-ent` and `?d1` bound to `pushed-dir`, the following statements will be inferred in the scenario:

```
(active Push05-Instance)

(causes
  (active Push05-Instance)
  (exists ?M1
    (and (isa ?M1 MovementEvent)
          (objectMoving ?M1 pushed-ent)
          (motionPathway ?M1 pushed-dir))))
```

The causal relation therefore indicates that the activation of the model fragment instance will cause a new `MovementEvent` with the pushed object moving in the direction of the push. This means that if this model is instantiated in a scenario, the agent should predict movement to occur as an effect, either in the present state or in a subsequent state.

Suppose, contrary to Figure 29, that `?m1` is actually a cause in the model rather than an effect. In this case, the following statement would be a consequence of the model:

```
(causes
  (exists ?M1
    (and (isa ?M1 MovementEvent)
          (objectMoving ?M1 ?P1)
          (motionPathway ?M1 ?D1)))
  (active ?self))
```

This states that if the agent must explain what caused the state of events represented in the constraints of the model, a `MovementEvent ?m1` is the cause. This means that if the model is instantiated in a scenario, the agent should predict some event `?m1` to also occur in the present state or to have occurred in the immediately preceding state. The presence of this causal factor within the consequences block of the model may seem counterintuitive, but we must not conflate logical consequences (e.g., as in model formulation) with causal consequences (i.e., effects).

For this simulation, we gave the system 17 multimodal comic graphs as training data. These comic graphs described 50 instances of events, all either `PushingAnObject`, `Blocking`, or `MovementEvent`. This resulted in 50 temporally-encoded microtheories describing each event instance, which resulted in ten SAGE generalizations (shown in Figure 27) after analogical learning. These were transformed into descriptive qualitative models of pushing (e.g., Figure 29), moving, and blocking, using the processes described above.

To summarize, SAGE generalizations are probabilistic abstractions of observations, but they are not causal models in themselves. These are converted into qualitative models in two steps: (1) filtering is used to select the high probability statements, and (2) using the temporal relations of these statements to determine their role in a qualitative model. We next discuss how these qualitative models compare to the mental models of students on two problem-solving tasks.

#### **5.4 Comparing the system's models of motion to students' mental models**

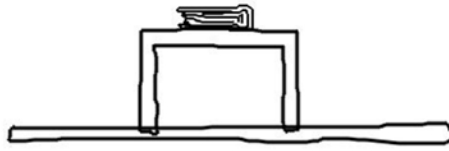
We cannot directly observe students' mental models – if we could, there would be little question of how they are represented and how they change. Consequently, we can only compare the system's models to students' mental models by comparing the predictions and explanations they generate during problem-solving tasks. We chose two problems from the learning science

literature: one from Brown (1994), and one from the *Force Concept Inventory* (Hestenes et al., 1992). We discuss each problem, the results from students, and the results from our simulation.

Brown (1994) showed a group of 73 high-school students a book resting on the surface of a table, and asked them whether the table exerts a force on the book. Here are the most popular answers provided by the students:

1. *Yes*. The table must exert an upward force the book to counteract the downward force of the book (33 students).
2. *No*. Gravity pushes the book flat, and the book exerts a force on the table. The table merely supports the book (19 students).
3. *No*. The table requires energy to push (7 students).
4. *No*. The table is not pushing or pulling (5 students).
5. *No*. The table is just blocking the book (4 students).
6. *No*. The book would move upward if the table exerted a force (4 students).

Thirty-three students correctly explained that the table pushes up against the book. The forty-student majority denied that the table exerted a force on the book, but for five different reasons (answers 2-6). Some students gave more than one incorrect explanation. For our present purposes, we are interested in modeling the incorrect answers, because these are the intuitive models of dynamics that students hold prior to conceptual change. If the simulation's qualitative models are comparable in content to student mental models, then the simulation will make the same set of mistakes as students.



**Figure 30: The sketch for the problem-solving task from Brown (1994).**

Our simulation was given a sketch of the same problem, illustrated in Figure 30. The system had a domain theory containing the qualitative models learned via SAGE and the facts that the omnipresent force of gravity pushes all things (i.e., instances of `Entity`) downward, but is not an `Entity` itself. Given the sketched scenario illustrated in Figure 30, we queried the system to (1) find all instances of `PushingAnObject` that are consistent with the scenario and then (2) explain why a `PushingAnObject` event between the table and the book in the upward direction must or must not exist.

To complete the first task, the system uses model-based inference (described in Chapter 3) to instantiate all qualitative models whose participants and constraints are satisfied in the scenario. Specifically, the system begins by inferring that gravity pushes all objects downward and then instantiates its qualitative models to create causal explanations and predictions about these `PushingAnObject` events. All of these events were explained by a model that relates `PushingAnObject` and `Blocking`. This model was used to infer two blocking events:

1. Gravity *pushes* down on the book which *pushes* down on the table, and the table *blocks* the book.
2. Gravity *pushes* down on the table which *pushes* down on the ground, and the table *blocks* the ground.

The first inference is similar to student answers (2) and (5) above, used a total of 23 times in Brown's (1994) experiment, except the simulation does not mention the concept of *support* in student answer (2). This explanation given by the students and the system does not directly confirm or deny that the table pushes the book, but it does describe the causal relationship between pushing and blocking within the scenario.

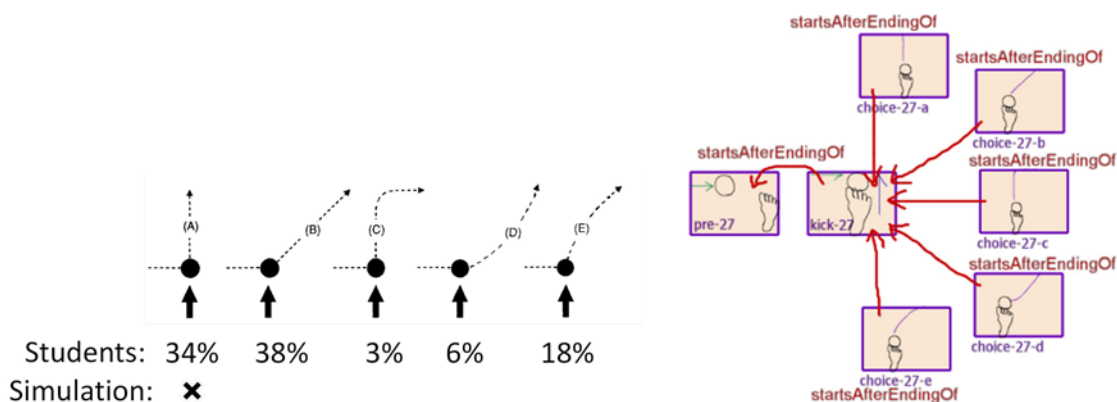
The system's second task is to explain why the table must or must not push the book, if there is sufficient evidence present. This involves (1) assuming that the `PushingAnObject` does in fact occur in the scenario, (2) instantiating qualitative models as in the first task, and then (3) searching for contradictions that arise as a result. Contradictions are found by querying for inconsistent patterns, e.g., a statement and its negation are simultaneously believed in the same state or an observable event (e.g., `MovementEvent`) is inferred but not observed in the scenario. If a contradiction is found, this constitutes an indirect proof that the table does not push the book. The system uses the qualitative model in Figure 29 to achieve this. This results in the following inference:

3. The table pushing the book would result in the book moving upward.

Since movement is not observed, this is contradictory.

This inference is similar to student answer (5), used by four students in Brown (1994).

In one multiple choice question from the Force Concept Inventory (Hestenes et al., 1992), students are shown a top-down sketch of a puck sliding to the right along a frictionless surface, and asked which path it would traverse if given an instantaneous kick forward. The problem and the proportion of student responses are shown in Figure 31, left. We sketched the problem using CogSketch as a comic graph with a fork in the state space (Figure 31, right), such that after the kick, the puck could traverse one of five different paths (a-e). The simulation decides which path the puck will traverse by exhaustively instantiating models qualitative models in the pre-fork state where the foot kicks the puck. The answer (a-e) that matches the simulation's prediction is chosen.



**Figure 31: Problem from the Force Concept Inventory, and student/simulation responses (left). Sketch of the same problem using CogSketch (right).**

The only model that can be instantiated (i.e., its participants and conditions are satisfied) in this scenario is the model in Figure 29. The causal consequence of this instance is that the puck (bound to slot ?p2) is the subject of a `MovementEvent` in the direction ?d1 (bound to Up). This behavior is described in choice (a) in the scenario, so this is the choice made by the system. Choice (a) was the most popular incorrect answer of the students tested by Hestenes et al. (1992), which suggests that it is a common misconception.

## 5.5 Discussion

This simulation induces descriptive qualitative models from observations using analogical generalization. When the qualitative models were used to solve two problems from the learning science literature, they produced some of the same incorrect explanations and predictions as novice students.

The fact that our system uses qualitative models to simulate some of the predictions and explanations of novices supports the claim that qualitative models provide a consistent computational account of human mental models. Since these models were induced from sketched observations, this simulation also supports the claim that analogical generalization is capable of inducing qualitative models. Importantly, the qualitative models learned by this simulation are mechanism-free, since they only describe causal relationships between discrete events. Since novices and experts alike are capable of explaining mechanisms of change (e.g., physical processes and influences between quantities), more evidence is needed to support the first claim.

The match between our system and novice students rely upon the psychological assumptions of our model discussed in Chapter 1 and the perceptual assumptions about comic graphs discussed in Chapter 3. To summarize: the training data of this simulation are sparser than the observations human encounter in the world because they contain only causally-relevant entities (e.g., there are no birds flying overhead) and they are already segmented into meaningful qualitative states. These simplifications reduce the complexity of learning and permit the system to learn much faster than people. Further, since the system has complete information about each state, each event (e.g., instance of `PushingAnObject`) is always observed in conjunction with its constraints (e.g., `touching` statements). This means that there is a perfect correlation for



events and their observed constraints in this simulation, but this information is not always available to people.

We have sketched a simplified account of how people might develop mental models from observing the world: abstracting common structure and inferring causal relations based on temporal relations. Whether the system can learn scientifically-accurate, Newtonian models via observation is an empirical question. It is not a question of knowledge representation, since qualitative models can represent scientifically-accurate models of dynamics (see Forbus, 1984); rather, it is a question of the inductive learning process. And since the vast majority of students only develop a Newtonian understanding of the world after formal instruction, we should not expect an accurate model of human learning to induce Newtonian dynamics from observation alone.

This first simulation only utilizes the explanation-based network inasmuch as it generates qualitative models to populate the domain theory, and then uses these models to make inferences during problem solving. The next simulations address how qualitative models change, provided instruction and interaction.

## Chapter 6: Revising mechanism-based models of the seasons

Thus far, we have simulated how mental models might be induced from observations. However, this does not account for how mental models change, or how knowledge is incorporated via communication or instruction. The simulation in this chapter addresses these two topics. We model middle-school students in a study by Sherin et al. (2012) who construct – and in some cases, revise – explanations of why the seasons change, during a clinical interview.

This simulation and the simulations in Chapters 7-8 assume that when a student revises her mechanism-based explanation of a phenomenon, such as seasonal change, she has also revised her underlying mental model of that phenomenon. Recall that in Chapter 1, we assumed that mental models are used to construct explanations of phenomena. If a student revises her explanation, she has constructed a new explanation that she prefers over her previous explanation. More specifically, she has recombined her knowledge into a different mental model, and its structure, assumptions, and inferences are preferable to that of the former mental model in the context of the phenomenon explained. So, explanation revision is a good indicator of mental model revision.

This simulation provides support for claims 1 and 3 of this dissertation:

***Claim 1:*** Compositional qualitative models provide a consistent computational account of human mental models.

***Claim 3:*** Human mental model transformation and category revision can both be modeled by iteratively (1) constructing explanations and (2) using meta-level reasoning to select among competing explanations and revise domain knowledge.

Since we are simulating how students reason about seasonal change, we represent student domain knowledge with qualitative model fragments to support claim 1. We use the explanation-based model of conceptual change described in Chapter 4 to simulate mental model transformation and support claim 3. We begin by discussing the learning science study with students, and then we discuss our simulation setup.<sup>32</sup>

### 6.1 How commonsense explanations (and seasons) change

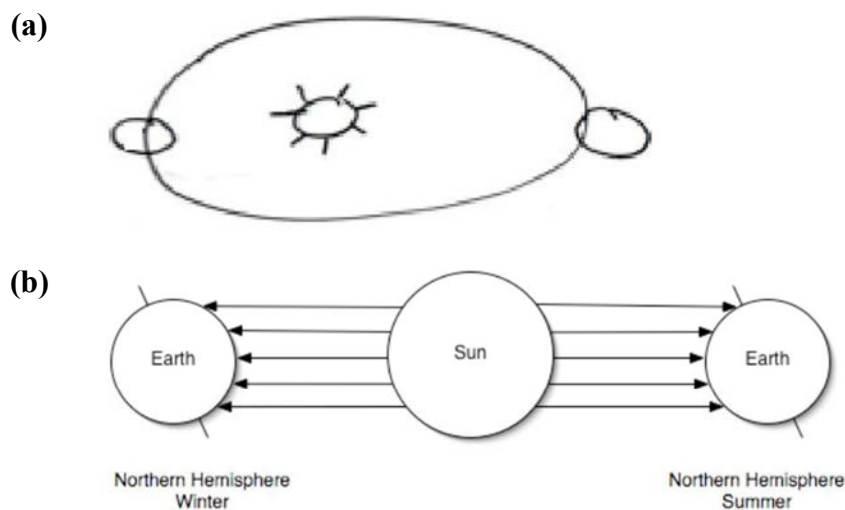
The experimenters in Sherin et al. (2012) interviewed 35 middle-school students regarding the changing of the seasons to investigate how students use commonsense science knowledge. Each interview began with the question “Why is it warmer in the summer and colder in the winter?” followed by additional questions and sketching for clarification. If the interviewee’s initial explanation of seasonal change did not account for different parts of the earth experiencing different seasons simultaneously, the interviewer asked, “Have you heard that when it’s summer [in Chicago], it is winter in Australia?” This additional information, whether familiar or not to the student, often lead the student to identify an inconsistency in their explanation and reformulate an answer to the initial question by recombining existing beliefs.

The interview transcript from the student “Angela” is listed in the appendix, courtesy of Sherin et al. Angela begins by explaining that the earth is closer to the sun in the summer than in the winter, which we call the *near-far* explanation. The seasons change as the earth approaches and retreats from the sun throughout its orbit of the sun. This is illustrated by a student sketch in Figure 32a. When the interviewer asks Angela if she has heard that Australia experiences its

---

<sup>32</sup> This builds upon the simulation published in Friedman, Forbus, & Sherin (2011)

winter during Chicago's summer, and whether this is a problem for her explanation, Angela sees that her explanation is problematic. She eventually changes her answer by explaining that the spin of the earth changes the seasons: the parts of the earth that face the sun experience their summer, while the parts that face away experience winter. We call this the *facing* explanation. Other students used the near-far explanation and the facing explanation, and many students drew upon idiosyncratic knowledge, e.g., that they had seen a picture of a sunny day in Antarctica, which influenced their explanations.



**Figure 32: Two diagrams explaining seasonal change, courtesy of Sherin et al. (2012). (a) Sketch from a novice student, explaining that the earth is closer to the sun in the summer than in the winter. (b) Scientific explanation involving tilt and insolation.**

The interviewer did not relate the correct scientific explanation during the course of the interview, so the students transitioned between various intuitive explanations. The scientifically accurate explanation of the seasons is that the earth's axis of rotation is tilted relative to its orbital plane, so it always points in the same direction throughout its orbit around the sun. When the northern hemisphere is inclined toward the sun, it receives more direct sunlight than when tilted away, which results in warmer and cooler temperatures, respectively. This is illustrated in

Figure 32b. While many students mentioned that the earth's axis is tilted, fewer used this fact in an explanation, and none of these were scientifically accurate.

Sherin et al. created a master listing of conceptual knowledge used by the students during the interviews, including propositional beliefs, general schemas, and fragmentary mental models. Five of the students from the study were characterized with enough precision for us to encode their beliefs and mental models using propositions and qualitative model fragments, respectively.

The rest of this chapter describes a simulation of how these five students construct explanations of dynamic systems from fragmentary domain knowledge and how these explanations are revised after new information renders them inconsistent. Each trial of the simulation corresponds to a subset of these students, so the starting domain knowledge varies across the trials, but the rest of the simulation is identical. We use the Angela trial to describe the workings of the simulation. As mentioned above, the students interviewed were not given the correct explanation, but we include an additional simulation trial that has access to the knowledge required for the correct explanation. This demonstrates that the system can construct the correct explanation when provided correct domain knowledge.

Our simulation of the students in Sherin et al. uses the conceptual change model described in Chapter 4 including: (1) the explanation-based network; (2) qualitative model fragments; (3) the abductive model formulation algorithm; and (4) cost functions to compute preferences over explanations. We next describe how these processes construct and revise qualitative models and explanations.

```

ModelFragment AstronomicalHeating
Participants:
  ?heater HeatSource (providerOf)
  ?heated AstronomicalBody (consumerOf)
Constraints:
  (spatiallyDisjoint ?heater ?heated)
Conditions: nil
Consequences:
  (qprop- (Temp ?heated) (Dist ?heater ?heated))
  (qprop (Temp ?heated) (Temp ?heater))

ModelFragment Approaching-PeriodicPath
Participants:
  ?mover AstronomicalBody (objTranslating)
  ?static AstronomicalBody (to-Generic)
  ?path Path-Cyclic (alongPath)
  ?movement Translation-Periodic (translation)
  ?near-pt ProximalPoint (toLocation)
  ?far-pt DistalPoint (fromLocation)
Constraints:
  (spatiallyDisjoint ?mover ?static)
  (not (centeredOn ?path ?static))
  (objectTranslating ?movement ?mover)
  (alongPath ?movement ?path)
  (on-Physical ?far-pt ?path)
  (on-Physical ?near-pt ?path)
  (to-Generic ?far-pt ?static)
  (to-Generic ?near-pt ?static)
Conditions:
  (active ?movement)
  (betweenOnPath ?mover ?far-pt ?near-pt)
Consequences:
  (i- (Dist ?static ?mover) (Rate ?self))

```

When an astronomical body *heated* and a heat source *heater* are spatially separated, the temperature of *heated*: (1) increases with the temperature of *heater* and (2) decreases as the distance between them increases.

An object *mover* travels on a cyclic path *path* relative to another object *static* where *path* is not centered on *static*. If *mover* is approaching – but not at – the closest point on *path* to *static*, then there is a rate of approach which decreases the distance from *mover* to *static*.

**Figure 33: AstronomicalHeating (top) and Approaching-PeriodicPath (bottom) model fragments used in the simulation. English interpretations of both model fragments included at right.**

## 6.2 Simulating how students construct and revise explanations

The students interviewed by Sherin et al. (2012) performed two tasks that are especially relevant to conceptual change.

1. Explain existing beliefs (e.g., Chicago and Australia are warmer in their summers than they are in their winters) when prompted.
2. Incorporate new, credible, information (e.g., Chicago's summer coincides with Australia's winter) and change explanations as needed to improve coherence.

These are the tasks we are interested in simulating in this chapter. We model the first task by (1) using the abductive model formulation algorithm described in Chapter 4 to construct explanations and then (2) using the cost function to determine which explanation is preferred. We model the second task by (1) adding new domain knowledge, (2) searching for contradictions, and then (3) using the cost reduction procedure described in Chapter 4 (*restructure-around-artifact* in Figure 23) to find more suitable sets of explanations for existing beliefs, when possible.

For the Angela trial, the system starts with a set of model fragments for both the near-far explanation and the facing explanation, since Angela constructed both of these explanations during the interview without learning these concepts from the interviewer. Two of these model fragments and their simplified English translations are shown in Figure 33. The system also contains propositional beliefs, such as the belief that Chicago is warmer in its summer than in its winter. This belief is represented by the following statement:

```
(greaterThan (M (Temp Chicago) ChiSummer)
              (M (Temp Chicago) ChiWinter))
```

The `M` function in this statement take two arguments – a quantity term such as `(Temp Chicago)` and a state such as `ChiSummer` – and denotes the measurement of the quantity within the state. This statement therefore translates to “the temperature of Chicago is greater in its summer than in its winter.” `ChiSummer` and `ChiWinter` are the subjects of other beliefs in the system’s domain knowledge beliefs such as:

```
(isa ChiWinter CalendarSeason)
(isa ChiAutumn CalendarSeason)
(isa ChiSummer CalendarSeason)
(isa ChiSpring CalendarSeason)
(contiguousAfter ChiWinter ChiAutumn)
(contiguousAfter ChiAutumn ChiSummer)
(contiguousAfter ChiSummer ChiSpring)
(contiguousAfter ChiSpring ChiWinter)
```

These beliefs, including the `greaterThan` statement, are all present in the system’s adopted domain knowledge microtheory  $\mathbb{D}_a$  at the beginning of the simulation trial, but they are not yet used within any explanations.

### 6.2.1 Explaining Chicago’s seasons

At the beginning of our Angela trial, we query the system for an explanation of why it is warmer in Chicago’s summer than in its winter. This is done by calling *justify-explanandum* in (Figure



34; but also see Chapter 4) with the following inputs: the `greaterThan` statement as the explanandum; the model fragments in  $\mathbb{D}_a$  as the domain theory; and the adopted domain knowledge microtheory  $\mathbb{D}_a$  as the scenario. The *justify-explanandum* procedure uses the abductive model formulation procedure (Chapter 4, Figure 17) to instantiate model fragments that help justify the explanandum. These procedures build the network structure for explanation  $x_I$  of Chicago's seasons shown in Figure 35. We step through the procedures in Figure 34 in greater detail to show how the explanation  $x_I$  in Figure 35 is constructed. Chapter 4 provided a detailed example of abductive model formulation, so we concentrate here on the *justify-ordinal-relation* and *justify-quantity-change* procedures that invoke the abductive model formulation procedure. We assume that an explanandum is one of the following: (1) a symbol that refers to a process instance; (2) an ordinal relation represented by a `greaterThan` statement; or (3) a quantity change represented by an `increasing` or `decreasing` statement. This means that our system does not justify the existence of physical objects, since we are primarily concerned with explaining physical phenomena and events. Also, our system does not justify `equalTo` relations, since – without information to the contrary – these can be explained by the absence of direct and indirect influences. Any `lessThan` relation can be converted into a `greaterThan` relation by reversing its two arguments.

When *justify-explanandum* is called on the belief that Chicago is warmer in its summer than its winter, the system detects that the explanandum is an ordinal relation, and invokes *justify-ordinal-relation*. This procedure binds  $q$  to  $(\text{Temp Chicago})$ ,  $s_1$  to  $\text{ChiSummer}$ , and  $s_2$  to  $\text{ChiWinter}$ . It then queries to determine whether (1)  $\text{ChiWinter}$  is after  $\text{ChiSummer}$  and whether (2)  $\text{ChiSummer}$  is

---

#### Front-ends to abductive model formulation

---

```

procedure justify-explanandum(explanandum m, domain D, scenario S)
  if  $m$  is a symbol and  $m$  is an instance of collection  $C$  such that  $(\text{isa } C \text{ ModelFragment})$ :
    justify-process( $m$ ,  $D$ ,  $S$ )
  else if  $m$  unifies with  $(\text{greaterThan } ?x \text{ } ?y)$ :
    justify-ordinal-relation( $m$ ,  $D$ ,  $S$ )
  else if  $m$  unifies with  $(\text{increasing } ?x)$  or with  $(\text{decreasing } ?x)$ :
    let  $q, d = \text{quantity-of-change}(m)$ ,  $\text{direction-of-change}(m)$ 
    justify-quantity-change( $q, d, D, S$ )

procedure justify-ordinal-relation (ordinal relation m, domain D, scenario S)
  //  $m$  is of the form  $(\text{greaterThan } (M \text{ } q \text{ } s_1) \text{ } (M \text{ } q \text{ } s_2))$ 
  let  $q, s_1, s_2 = \text{quantity-of}(m)$ ,  $\text{state-1-of}(m)$ ,  $\text{state-2-of}(m)$ 
  if query  $S$  for  $(\text{after } s_2 \text{ } s_1)$  then:
    justify-quantity-change( $q, i-, D, S$ )
  if query  $S$  for  $(\text{after } s_1 \text{ } s_2)$  then:
    justify-quantity-change( $q, i+, D, S$ )

procedure justify-quantity-change (quantity q, direction d, domain D, scenario S)
  // Find direct and indirect influences of  $q$ 
  instantiate-fragments-with-consequence( $(\text{qprop } q \text{ } ?x)$ ,  $D, S$ )
  instantiate-fragments-with-consequence( $(\text{qprop- } q \text{ } ?x)$ ,  $D, S$ )
  instantiate-fragments-with-consequence( $(d \text{ } q \text{ } ?x)$ ,  $D, S$ )
  let  $I_i = \text{query } S \text{ for indirect influences on } q$ . // results are in form  $(\text{qprop/qprop- } q \text{ } ?x)$ 
  for each  $i$  in  $I_i$ :
    let  $d_i = \text{direction-of-influence}(i)$  //  $\text{qprop}$  or  $\text{qprop-}$ 
    let  $q_i = \text{influencing-quantity}(i)$ 
    let  $d_c = d$ 
    if  $d_i = \text{qprop-}$  then:
      set  $d_c = \text{opposite}(d)$ 
    justify-quantity-change( $q_i, d_c, D, S$ )

```

---

**Figure 34: Pseudo-code for constructing explanations about ordinal relations and quantity changes, from Chapter 4.**

after  $\text{ChiWinter}$ . Since both are true, the beliefs  $f_{19-20}$  in Figure 35 are encoded to justify the explanandum. Now the system must justify how  $(\text{Temp Chicago})$  decreases between  $\text{chiSummer}$

and `chiWinter` and how it increases between `chiWinter` and `chiSummer`. This is achieved with the following two procedure invocations:

*justify-quantity-change*(`(Temp Chicago)`, `i-`, `D`, `S`)

*justify-quantity-change*(`(Temp Chicago)`, `i+`, `D`, `S`)

Notice that these invocations make no mention of `ChiWinter` and `ChiSummer`. This is because the system is building a model of the mechanisms by which the temperature of Chicago might increase and decrease. These beliefs and causal mechanisms are explicitly quantified in specific states using temporal quantifiers represented as white triangles in Figure 35. We discuss temporal quantifiers before continuing our walk-through.

Consider the temporal quantifier that justifies  $f_{20}$  with  $f_{18}$  in Figure 35. This states that we can believe  $f_{20}$  (i.e., `(holdsIn (Interval ChiSummer ChiWinter) (decreasing (Temp Chicago)))`) so long as the belief  $f_{18}$  (i.e., `(decreasing (Temp Chicago))`) and all beliefs justifying  $f_{18}$  hold within the state `(Interval ChiSummer ChiWinter)`. This compresses the explanation structure: without these temporal quantifiers, we would have to store each belief  $b$  left of  $f_{20}$  as `(holdsIn (Interval ChiSummer ChiWinter) b)`. The temporal quantifiers in the network can be used to decompress the explanation into this format without any loss of information, but we can perform temporal reasoning without decompressing.

The invocation of *justify-quantity-change*(`(Temp Chicago)`, `i-`, `D`, `S`) begins by abductively instantiating all model fragments in the domain theory that contain a consequence that unifies with `(qprop (Temp Chicago) ?x)`, `(qprop- (Temp Chicago) ?x)`, or `(i- (Temp Chicago) ?x)`. This uses the abductive model formulation algorithm described in Chapter 4. The result is the instantiation of qualitative models that can contain indirect (i.e., `qprop` and

qprop-) and direct (i.e., i-) influences on Chicago's temperature to help explain why it decreases. After these invocations, the procedure *justify-quantity-change* finds these and other influences which explain Chicago's temperature decreasing within the scenario model. In our example, it finds the qualitative proportionality (qprop (Temp Chicago) (Temp PlanetEarth)) represented as  $f_{16}$  in Figure 35, which states that the temperature of Chicago will decrease if the temperature of the earth decreases. Next the system attempts to justify the earth decreasing in temperature (decreasing (Temp PlanetEarth)), plotted as  $f_{14}$  in Figure 35. This results in the recursive invocation:

*justify-quantity-change*((Temp PlanetEarth), i-,  $D, S$ )

In this recursive invocation, the system finds the model fragment `AstronomicalHeating` (shown in Figure 33) with the following consequences:

```
(qprop- (Temp ?heated) (Dist ?heater ?heated))
(qprop (Temp ?heated) (Temp ?heater))
```

When the system binds `?heated` to `PlanetEarth` and invokes abductive model formulation, it instantiates and activates an instance of `AstronomicalHeating` with produces the statements  $f_{9-11}$  in Figure 35, including:

```
(qprop- (Temp PlanetEarth) (Dist TheSun PlanetEarth))
(qprop (Temp PlanetEarth) (Temp TheSun))
```

Consequently, when the procedure next searches for indirect influences of (Temp PlanetEarth), it determines that it can justify the earth's cooling with an increase of (Dist TheSun PlanetEarth) or a decrease of (Temp TheSun). This makes another recursive invocation of *justify-quantity-change* to justify an increase of (Dist TheSun PlanetEarth). This ultimately creates a Retreating-Periodic instance whose rate increases the earth's distance to the sun (statement  $f_{12}$  in Figure 35) during a segment of the earth's orbit around the sun.

### Legend

$f_0$	(isa earthPath EllipticalPath)	$f_9$	(active AH-inst)
$f_1$	(spatiallyDisjoint earthPath TheSun)	$f_{10}$	(qprop- (Temp PlanetEarth) (Dist TheSun PlanetEarth))
$f_2$	(isa TheSun AstronomicalBody)	$f_{11}$	(qprop (Temp PlanetEarth) (Temp TheSun))
$m_0$	(isa ProximalPoint ModelFragment)	$f_{12}$	(i+ (Dist TheSun PlanetEarth) (Rate RPP-inst))
$m_1$	(isa DistalPoint ModelFragment)	$f_{13}$	(increasing (Temp PlanetEarth))
$m_2$	(isa Approaching-Periodic ModelFragment)	$f_{14}$	(decreasing (Temp PlanetEarth))
$m_3$	(isa AstronomicalHeating ModelFragment)	$f_{15}$	(qprop (Temp Australia) (Temp PlanetEarth))
$m_4$	(isa Retreating-Periodic ModelFragment)	$f_{16}$	(qprop (Temp Chicago) (Temp PlanetEarth))
$f_3$	(isa TheSun HeatSource)	$f_{17}$	(increasing (Temp Chicago))
$f_4$	(spatiallyDisjoint TheSun PlanetEarth)	$f_{18}$	(decreasing (Temp Chicago))
$f_5$	(isa APP-inst Approaching-PeriodicPath)	$f_{19}$	(holdsIn (Interval ChiWinter ChiSummer) (increasing (Temp Chicago)))
$f_6$	(isa AH-inst AstronomicalHeating)	$f_{20}$	(holdsIn (Interval ChiSummer ChiWinter) (decreasing (Temp Chicago)))
$f_7$	(isa RPP-inst Retreating-PeriodicPath)	$f_{21}$	(greaterThan (M (Temp Australia) AusSummer) (M (Temp Australia) AusWinter))
$f_8$	(i- (Dist TheSun PlanetEarth) (Rate APP-inst))	$f_{22}$	(greaterThan (M (Temp Chicago) ChiSummer) (M (Temp Chicago) ChiWinter))

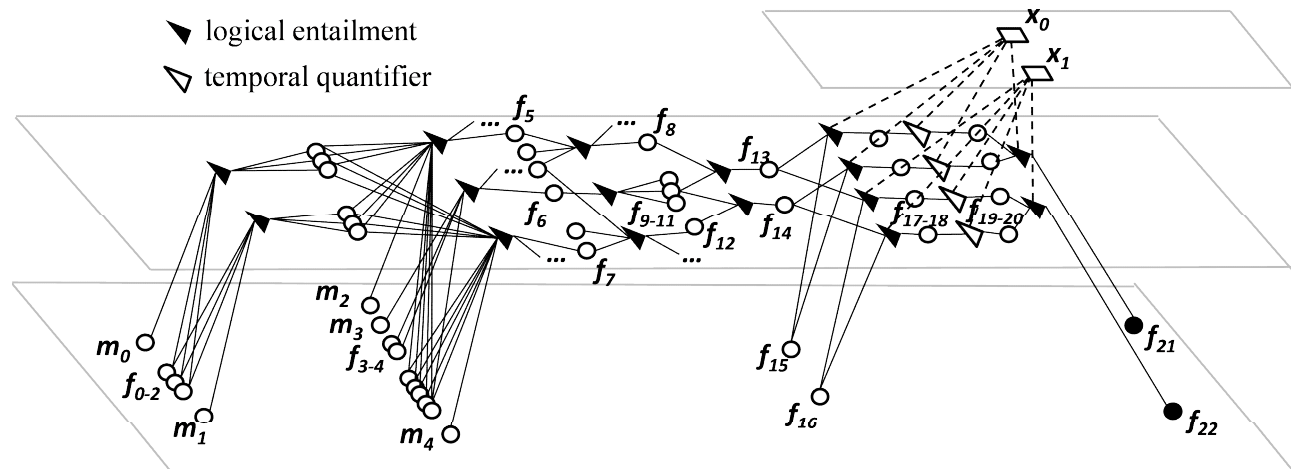


Figure 35: Network plotting explanations  $x_0$  and  $x_1$  that explain seasonal change in Australia ( $x_0$ ) and Chicago ( $x_1$ ) using a near-far model of the seasons.

We have described how the system justifies Chicago decreasing in temperature. The system justifies Chicago's *increase* in temperature in an analogous fashion. It uses some of the model fragment instances created to explain Chicago's decrease in temperature, such as the `AstronomicalHeating` instance. It also instantiates new model fragments, such as an `Approaching-Periodic` instance whose rate decreases the earth's distance to the sun (statement  $f_8$  in Figure 35) which justifies the earth's increase in temperature (statement  $f_{13}$  in Figure 35).

After the system has computed the justification structure for the explanandum, it finds all well-founded explanations of the explanandum and creates a unique explanation node (e.g.,  $x_I$  in Figure 35) for each. As we discussed in Chapter 4, multiple explanations may compete to explain the same explanandum. In our simulation of Angela, there are multiple explanations for Chicago's seasons, only one of which ( $x_I$ ) is shown in Figure 35. Consider the following simplified explanations in English:

- $x_I$ : The earth retreats from the sun for Chicago's winter and approaches for its summer (shown in Figure 35).
- $x'_I$ : The sun's temperature decreases for Chicago's winter and increases for its summer.
- $x'_2$ : The sun's temperature decreases for Chicago's winter, and the earth approaches the sun for its summer.
- $x'_3$ : The earth retreats from the sun for Chicago's winter, and the sun's temperature increases for its summer.

Explanations  $\{x_I, x'_I, x'_2, x'_3\}$  compete with each other to explain  $f_{22}$ . However,  $x'_I, x'_2$ , and  $x'_3$  are all problematic. Explanations  $x'_2$  and  $x'_3$  contain asymmetric quantity changes in a cyclic state space: a quantity (e.g., the sun's temperature) changes in the summer  $\rightarrow$  winter interval without returning to its prior value somewhere in the remainder of the state cycle, winter  $\rightarrow$  summer. Explanation  $x'_I$  is not structurally or temporally problematic, but the domain theory contains no model fragments that can describe the process of the sun changing its temperature. Consequently, the changes in the sun's temperature are *assumed* rather than justified by process instances. Assumed quantity changes are problematic because they represent unexplainable changes in a system. These are also problematic under the *sole mechanism assumption* (Forbus, 1984), which states that all changes in a physical system are the result of processes.<sup>33</sup> We have just analyzed and discredited system-generated explanations  $x'_{I-3}$  which compete with explanation  $x_I$ . The system makes these judgments automatically, using the artifact-based cost function described in Chapter 4.

The cost function computes the cost of an explanation as the sum of the cost of new artifacts (e.g., model fragments, model fragment instances, assumptions, contradictions, etc.<sup>34</sup>) within that explanation. In our example,  $x_I$  is the preferred (i.e., lowest cost) explanation, so the system assigns  $x_I$  to the explanandum within the preferred explanation mapping  $\mathbb{E}$ , and thereby explains Chicago's temperature variation using the near-far model.

---

<sup>33</sup> The agent might explicitly assume that an unknown, active, process is directly influencing the quantity, but such an assumption is still objectively undesirable within an explanation.

<sup>34</sup> For a complete listing of epistemic artifacts and their numerical costs, see section 4.6.2.

### 6.2.2 Explaining Australia's seasons

We next query the system for an explanation of why Australia is warmer in its summer than in its winter. This invokes *justify-explanandum* which constructs explanations for Australia's seasons, including the explanation  $x_0$  in Figure 35. When the system chooses among competing explanations for Australia's seasons using the cost function, the cost of each explanation is influenced by the explanations it has chosen for previous explanandums (e.g., Chicago's seasons). This is because artifacts only incur a cost if they are not presently used in a preferred explanation. All else being equal, the system is biased to reuse existing artifacts such as model fragments (e.g., `AstronomicalHeating`), model fragment instances (e.g., `AstronomicalHeating` instance `AH-inst` represented as  $f_6$  in Figure 35), and assumptions that are in other preferred explanations. This causes the system to choose a near-far explanation for Australia's seasons ( $x_0$  in Figure 35) which contains much of the justification structure of the preferred explanation for Chicago's seasons ( $x_1$  in Figure 35).

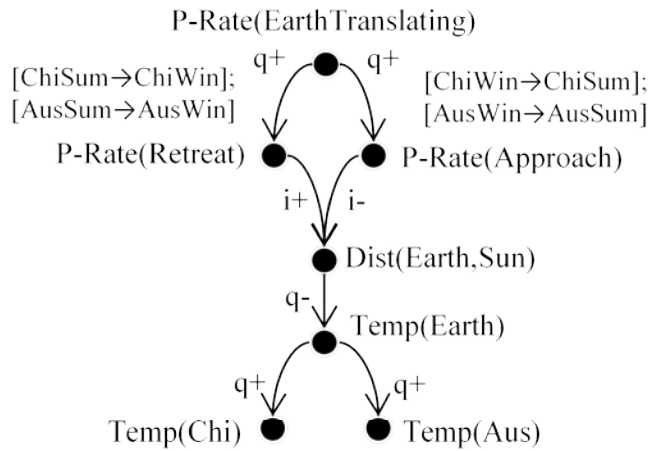
### 6.2.3 Comparing the system's explanations to student explanations

At this point, we want the system to describe the mechanisms that cause seasonal change and temperature change. Sherin et al. do not give the interviewees a pretest or posttest; rather, they ask the student to explain it freely. Generating causal explanations in English is outside the scope of this research, so we have our system describe causal models using *influence graphs* as illustrated in Figure 36. Given one or more explanations, the system automatically constructs an influence graph of the explanations by (1) creating a vertex for every quantity described in the explanation and (2) creating a directed edge for every influence described in the explanation. In



the case of Figure 36, the system graphs the two preferred explanations, so that both Australia's seasons and Chicago's seasons are explainable using the same mechanisms.

The majority of the influence graph in Figure 36 describes continuous causal mechanisms that are common to both explanations. The only explanation-specific components are the temperatures of Chicago and Australia and their qualitative proportionalities to the temperature



**Figure 36: An influence diagram of the near-far explanation of both Chicago's (Chi) and Australia's (Aus) seasons. Nodes are quantities and edges describe positive and negative direct influences (i+, i-) and indirect influences (q+, q-). Bracketed ranges quantify process activity.**

of the earth. This illustrates how knowledge is reused across explanations and how new phenomena are explained in terms of existing causal structure. Thus, even though explanations exist as separate entities in our computational model, they share significant structure.

#### 6.2.4 Accommodating new, credible information

Thus far, we have described how the system constructs and computes preferences for the two explanations plotted in Figure 35: one for how Chicago's seasons change ( $x_1$ ) and another for how Australia's seasons change ( $x_0$ ). Other explanations for Chicago's and Australia's seasons exist in the system, but are not preferred since they incur a greater cost.

In Sherin et al.’s study, recall that if a student’s explanation did not account for different seasons in different parts on the earth – like our simulation’s presently-preferred explanations – the interviewer asked them whether they were aware that Chicago’s winter coincided with Australia’s summer. This caused some students, including Angela, to revise their explanation of seasonal change. This section describes how we simulate the incorporation of new information and the subsequent explanation revision.

To begin, the following statements are added from the human user:

```
(cotemporal ChiSummer AusWinter)
(cotemporal ChiAutumn AusSpring)
(cotemporal ChiWinter AusSummer)
(cotemporal ChiSpring AusAutmn)
```

We refer to this as the *opposite seasons information*. These statements are from a trusted source, so each statement incurs a credibility artifact<sup>35</sup> of cost -1000 (where negative cost indicates a benefit). This means that for each of these four statements, the system receives a numerical benefit as long as it keeps the statement in the adopted domain knowledge microtheory  $\mathbb{D}_a$ . It will lose this benefit if it removes the statement from  $\mathbb{D}_a$ , though the statement will continue to exist in the general domain knowledge microtheory  $\mathbb{D}$ .

After adding these statements to  $\mathbb{D}_a$  the system searches for contradictions across its preferred explanations (i.e.,  $x_0$  and  $x_1$  in Figure 35) and adopted domain knowledge in  $\mathbb{D}_a$ . This is performed with domain-general rules for detecting contradictions, such as:

---

<sup>35</sup> See the section 4.6.2 for an overview and example of credibility artifacts.

- A belief and its negation cannot be simultaneously believed.
- A quantity cannot simultaneously be greater than  $n$  and less than or equal to  $n$ .
- A quantity cannot simultaneously be less than  $n$  and greater than or equal to  $n$ .

The quantity rules also apply to derivatives of quantities, so the system detects when quantities are believed to simultaneously increase and decrease.

To illustrate this behavior within the Angela example, consider Australia's explanation  $x_0 = \langle J_0, B_0, M_0 \rangle$  and Chicago's explanation  $x_1 = \langle J_1, B_1, M_1 \rangle$ . According to the definition of explanations in Chapter 4,  $B_0$  is the set of beliefs in  $x_0$  and  $B_1$  is the set of beliefs in  $x_1$ . Since both explanations refer to the near-far model, the following statements (as well as many others) are included in these belief sets:

$B_0$  contains the temporally-quantified statement:

```
(holdsIn (Interval AusSummer AusWinter)
  (decreasing (Temp PlanetEarth)))
```

(i.e., "Between Australia's summer and its winter, the earth cools.")

$B_1$  contains the temporally-quantified statement:

```
(holdsIn (Interval ChiWinter ChiSummer)
  (increasing (Temp PlanetEarth)))
```

(i.e., "Between Chicago's winter and summer, the earth warms.")

Before the opposite seasons information was incorporated, these statements were not contradictory. After we add the opposite seasons information, the system infers that the interval from Australia's summer to its winter coincides with the interval from Chicago's winter to its

summer. Therefore, the earth's temperature is believed to increase and decrease simultaneously, which is an impossible behavior within a physical system. This is flagged by the contradiction detection rules, and the following contradiction artifact is created:

```
<Contra, { (cotemporal ChiSummer AusWinter),
            (cotemporal ChiWinter AusSummer),
            (holdsIn (Interval AusSummer AusWinter)
              (decreasing (Temp PlanetEarth))),
            (holdsIn (Interval ChiWinter ChiSummer)
              (increasing (Temp PlanetEarth))) }
```

Three additional contradictions are detected between these explanations: (1) the opposite simultaneous heating/cooling of the earth; (2) the earth simultaneously approaching and retreating for Chicago and Australia, respectively; and (3) the earth simultaneously retreating and approaching for Chicago and Australia, respectively. Artifacts are created for these contradictions as well. Each contradiction artifact incurs a cost of 100.

Despite gaining numerical benefits for accepting the instructional knowledge about opposite seasons in Chicago and Australia, the system has detected four contradictions and incurred the respective costs. Recall from Chapter 4 that the cost of an artifact, such as the contradiction artifact shown above, is only incurred if every constituent belief is either (1) in the adopted domain knowledge microtheory  $\mathbb{D}_a$  or (2) in the belief set of a preferred explanation in the explanation mapping  $\mathbb{E}$ . Consequently, these contradiction costs might be avoided – while still retaining the credibility benefits – by revising the  $\mathbb{E}$  or  $\mathbb{D}_a$ . This involves retracting beliefs from  $\mathbb{D}_a$  and switching its preferred explanation(s) to disable this contradiction artifact and other costly artifacts. This is the role of the procedure *restructure-around-artifact* described in Chapter 4 (Figure 23). When this procedure is called with one of the newly-incurred

contradiction artifacts as the input argument, the procedure finds (1) beliefs in  $\mathbb{D}_a$  that support the contradiction (i.e., the two `cotemporal` statements above) and (2) explanandums whose explanations support the contradiction (i.e., Chicago’s seasonal temperature difference and Australia’s seasonal temperature difference).

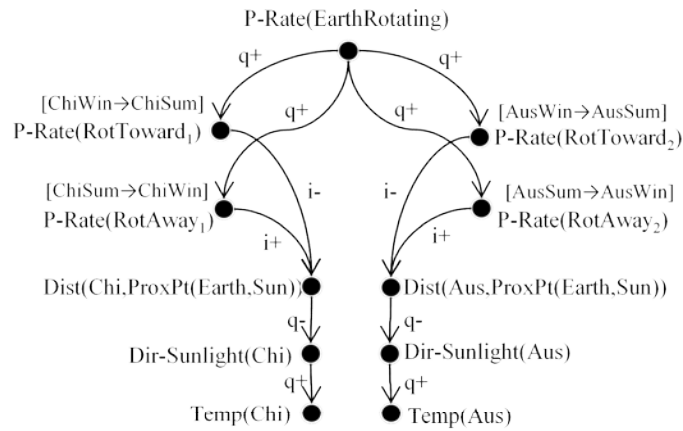
For each supporting belief in  $\mathbb{D}_a$ , the system determines whether removing the belief from  $\mathbb{D}_a$  will lower the overall cost. For example, removing `(cotemporal ChiSummer AusWinter)` from  $\mathbb{D}_a$  will remove all four contradictions for a benefit of 400, but it would also disable the credibility benefit of 1000, so there would be a net loss. Therefore, no change will be made here. The same is true of removing `(cotemporal ChiWinter AusSummer)` from  $\mathbb{D}_a$ .

For each supporting explanandum, the system computes the lowest cost explanation. For example, changing Chicago’s seasonal explanation to another explanation (e.g., the *facing* explanation, described above) revokes the beliefs that earth’s temperature and distance from the sun changes during Chicago’s seasonal intervals. The facing explanation was not initially the lowest-cost explanation for Chicago’s seasons, but these contradictions have since made the two near-far explanations much more costly.

When the system changes its explanation for Chicago’s seasons to the facing explanation, it disables all four contradictions; however, the ***restructure-around-artifact*** procedure is not yet complete. When it processes the final explanandum, Australia’s seasons, the system finds that it can further reduce cost by changing Australia’s preferred explanation from the near-far explanation to a facing explanation. This is because using the same model fragments, model fragment instances, and assumptions as Chicago’s new explanation (i.e., the facing model) is less expensive than using the near-far model to explain Australia’s seasons. The system then iterates

through the beliefs and explanandums again to determine whether additional unilateral changes can reduce cost, and since no further action reduces cost, the procedure terminates.

When the procedure terminates, both Chicago's and Australia's seasons have been mapped to explanations that use the facing model. The corresponding influence graph for both preferred explanations is shown in Figure 37. Both explanations use *RotatingToward* and *RotatingAway* processes to explain change in temperature, the rates of which are qualitatively proportional to the rate of the earth's rotation.



**Figure 37: An influence diagram of the facing explanation of both Chicago's (Chi) and Australia's (Aus) seasons.**

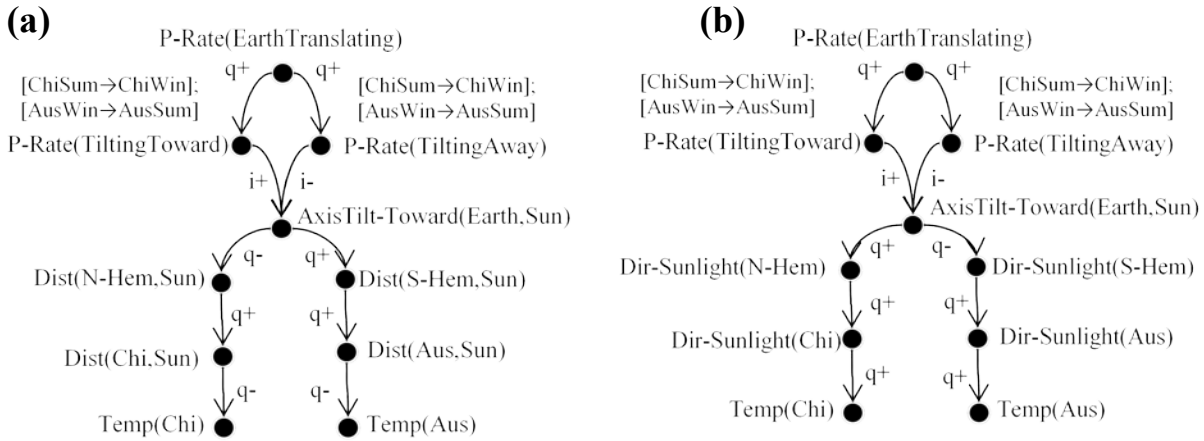
We have just described how the simulation accommodates new information by revising explanation preferences in  $\mathbb{E}$  to reduce cost. As we discussed in Chapter 4, the restructuring procedure is guaranteed to converge because it only performs belief revision if cost can be reduced, and cost cannot be reduced infinitely. Restructuring is a greedy algorithm, so it is not guaranteed to find the optimal cost configuration of explanation preferences.

This concludes the Angela trial. Like the student Angela, the computational model begins the session by explaining the seasons with a near-far explanation and ends the session with a

facing explanation. We simulate five of the students from Sherin et al.'s study, including Angela. We continue with a description of the simulation setup and experimental results.

### **6.3 Simulation results**

We implemented our model on top of the Companions cognitive architecture (Forbus et al., 2009), ran each trial as described above, and compared our system's explanations to those of students. In each trial, the system starts with a subset of knowledge pertaining to a student from Sherin et al., but no explanations have been constructed. In terms of Figure 35, the starting state of the system is a series of nodes on the bottom (domain theory) tier of the network, but none elsewhere. The system is then queried to construct explanations for Chicago's and Australia's seasons, after which we provide the simulation with the information about opposite seasons, and query the simulation again for an explanation of the seasons.



**Figure 38: Influence graphs for additional explanations produced by the simulation. (a) The tilt of the axis increases and decreases each hemisphere's distance to the sun. (b) A simplified correct explanation of the seasons.**

The individual differences of the students within the interviews involve more than just variations in domain knowledge. For example, some students strongly associate some models and beliefs with the seasons (e.g., that the earth's axis is tilted) without knowing the exact mechanism. To capture this (e.g., in the “Deidra & Angela” trial below), our system includes an additional numerical penalty over beliefs to bias explanation preference. We describe this further below.

**Ali & Kurt trial.** The system's initial domain knowledge includes: (1) the earth rotates on a tilted axis, (2) temperature is qualitatively proportional to sunlight, and (3) the earth orbits the sun. However, there is no knowledge that each hemisphere is tilted toward and away during the orbit. Consequently, the system computes nine explanations for both Chicago and Australia, and computes preference for the facing explanations shown in Figure 37, with a cost of 56. This explanation is consistent with the opposite seasons information, so no revision occurs as a result. Like Ali and Kurt, the simulation starts and ends with the facing explanation.



**Deidra & Angela trial.** The system's initial domain knowledge includes: (1) the earth rotates, (2) the earth orbits the sun and is sometimes closer and sometimes farther, and (3) sunlight and proximity to the sun both affect temperature. To model Deidra and Angela's preference for the distance-based explanation, for this trial we used an additional ten-point cost on the belief  $(qprop (Temp X) (Sunlight X))$ . Under these parameter settings, the system constructs 16 explanations<sup>36</sup> and computes a preference for the near-far explanations graphed in Figure 36, with a cost of 56. The system also created facing explanations (graphed in Figure 37) with a cost of 66, due to an additional ten-point penalty on the belief  $(qprop (Temp X) (Sunlight X))$ . This penalty makes the facing explanation more expensive than the near-far explanation. When confronted with the opposite seasons information, the system (like Deidra and Angela) detects inconsistencies and changes its preferred explanation from the near-far explanations to the facing explanations.

**Amanda trial.** The system's initial domain knowledge includes: (1) the earth orbits the sun, (2) the earth rotates on a tilted axis, (3) when each hemisphere is tilted toward the sun, it receives more sunlight and is more proximal to the sun, and (4) sunlight and proximity to the sun both affect temperature. In the interview, Amanda mentions two main influences on Chicago's temperature: (1) the distance to the sun due to the tilt of the earth, and (2) the amount of sunlight, also due to the tilt of the earth. Through the course of the interview, she settles on the latter. Amanda could not identify the mechanism by which the tilt changes throughout the year. We simulated Amanda once with process models for *TiltingToward*, and *TiltingAway*, producing graphs Figure 38(a) and Figure 38(b) with costs 52 and 67, respectively. However, since Amanda could not identify the processes that increased and decreased the tilt of the earth,

---

<sup>36</sup> The increased number of explanations is due to the belief that proximity in addition to amount of sunlight affect temperature.

we simulated her again without these process models. This produced two similar graphs, but without anything affecting `AxisTilt-Toward(Earth, Sun)`. This was the final model that the student Amanda chose as her explanation. The graphs in Figure 38 both describe the tilt of the earth as a factor of the seasons: graph (a) is incorrect because it describes tilt affecting distance and temperature, and graph (b) is a simplified correct model.

By varying the domain knowledge and manipulating the numerical costs of beliefs, we can use the simulation to (1) construct student explanations and (2) revise explanations under the same conditions as students. Further, in the Amanda trial, we provided additional process models to demonstrate that the simulation can construct a simplified correct explanation.

## 6.4 Discussion

In summary, this simulation (1) constructs explanations from available domain knowledge via abductive model formulation, (2) evaluates the resulting explanations using a cost function, and (3) detects inconsistencies and re-evaluates its explanations when given new information. By changing the initial knowledge of the system, we are able to simulate different interviewees' commonsense scientific reasoning regarding the changing of the seasons. We also demonstrated that the system can construct the scientifically correct explanation using the same knowledge representation and reasoning approaches.

This simulation supports the claim that model fragments can simulate mechanism-based psychological mental models. This is because model fragments (e.g., those in Figure 33) were used to describe processes and conceptual entities, and were able to capture the causal mechanisms of students' explanations. This simulation also supports the third claim of this dissertation: that conceptual change – in this case, mental model transformation – can be

simulated by constructing and evaluating explanations. The “Deidra & Angela” trial exemplifies this behavior by shifting explanations and underlying influence graphs (i.e., from that in Figure 36 to that in Figure 37), which represent different student mental models.

The numerical explanation scoring strategy used in this simulation is domain-general, albeit incomplete. To be sure, other factors not addressed by our cost function are also important considerations for explanation evaluation: belief probability, epistemic entrenchment, diversity of knowledge, level of specificity, familiarity, and the variable credibility of information (and information sources). Incorporating these factors will help model individual differences in response to instruction (e.g., Feltovich et al., 2001). We discuss some possible extensions in Chapter 9.

We believe that this simulation is doing much more computation than people to construct the same explanations. For example, the system computed and evaluated 16 explanations in the Deidra & Angela trial when explaining Chicago’s seasons. As described in Chapter 4, our system uses an abductive model formulation algorithm, followed by a complete meta-level analysis of competing explanations. People probably use a more incremental approach to explanation construction, where they interleave meta-level analysis within their model-building operations. Such an approach would avoid reifying explanations that are known to be problematic (e.g., explanations  $x'_{1-3}$  in section 6.2.1), but it would involve monitoring the model formulation process. The transcript of Angela’s interview in the appendix helps illustrate the incremental nature of psychological explanation construction: Angela appears to construct a second explanation only after she realizes that her initial explanation is flawed.

This simulation demonstrates that our computational model can reactively revise its explanations to maintain consistency and simplicity. However, this does not capture the entirety

of conceptual change, or even the entirety of mental model transformation. For instance, Angela and Deidra incorporated new information that forced them to recombine pre-existing knowledge into a new explanation, but they did not have to incorporate unfamiliar information about astronomy into their explanations. In contrast, when students learn from formal instruction or read from a textbook, they often encounter information about new entities, substances, and physical processes that must be incorporated into their current mental models. This is the subject of the simulation described in the next chapter.

## Chapter 7: Mental model transformation from textbook information

The last chapter simulated the revision of mechanism-based mental models when new information causes inconsistencies. Formal instruction can involve more subtle conflicts than this, such as learning about a biological system at a finer granularity and making sense of new entities and processes. Consider the mental model transformation example from Chapter 4: a student believes that blood flows from a single-chambered heart, through the human body, and back, and then reads that blood actually flows from the *left side* of the heart to the body. This new information does not directly contradict the student's mental model since the text does not explicitly state that blood does *not* flow from the heart; rather, the new information is more specific than the student's present mental model, and the conflict between beliefs and models is not as overt as it was in the previous chapter. This simulation constructs and evaluates explanations – similar to the previous chapter's simulation – to incrementally transform compositional qualitative models when provided a stream of textbook information. We simulate the students in Chi et al. (1994a) who complete a pretest about the circulatory system, read a textbook passage on the topic, and then complete a posttest to assess their learning.

Recall from Chapter 2 that act of explaining to oneself helps people revise flawed mental models (Chi et al., 1994a; Chi, 2000). Chi et al. (1994a) showed that when students are prompted to explain concepts to themselves while reading a textbook passage about the human circulatory system, they experience a greater gain in learning than students who read each sentence of the passage twice. Chi and colleagues call this the *self-explanation effect*. Chi (2000) describes how self-explanation causes mental model transformation:

1. Explaining the new information causes the recognition of qualitative conflicts (i.e., different predictions and structure) between the mental model and the model presented in the textbook.
2. The conflict is propagated in the mental model to find contradictions in the consequences.
3. The mental model is repaired using elementary addition, deletion, concatenation, or feature generalization operators.

The self-explanation effect is central to our computational model, but we do not implement it according to Chi's (2000) description. Our simulation simulates the psychological self-explanation effect by:

1. Constructing new explanations using new textbook information.
2. Evaluating the new explanations alongside previous ones.
3. Re-mapping explanandums to new explanations when preferences are computed as such.

As shown in the previous chapter's simulation, changing the preferred explanation for an explanandum can simulate belief revision. We describe this process in detail below.

Like the previous simulation, this simulation uses qualitative models to simulate students' mental models, and it uses the same central model of conceptual change. Consequently, this simulation provides additional support for the first and third claims of this dissertation:

**Claim 1:** Compositional qualitative models provide a psychologically plausible computational account of human mental models.

**Claim 3:** Human mental model transformation and category revision can both be modeled by iteratively (1) constructing explanations and (2) using meta-level reasoning to select among competing explanations and revise domain knowledge.

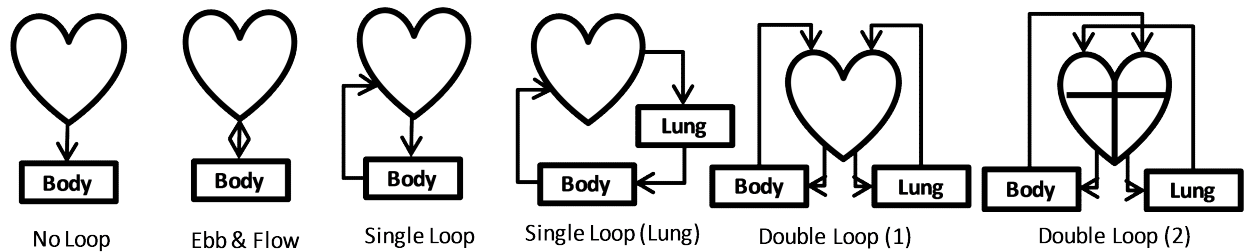
We briefly discuss Chi et al.'s (1994a) study, which is the basis for comparison in this simulation. We then discuss how textbook information is integrated into our system via explanation construction, and the results of our simulation (published as Friedman & Forbus, 2011).

### **7.1 Self-explaining improves student accommodation of textbook material**

Chi et al. (1994a) studied the self-explanation effect on 21 eighth-grade students. Each student was given a pretest to assess their knowledge of the human circulatory system. Each student then read a 101-sentence textbook passage about the circulatory system, after which they completed a posttest. There were two conditions: the control group (9 students) read each sentence in the passage twice, and the experimental group (12 students) read the passage once, but was prompted by the experimenter to explain portions of the text throughout the reading.

Part of the pretest and posttest involved plotting the flow of oxygen-rich and oxygen-poor blood through the human body, using arrows between various parts of the body. The tests also included conceptual questions about the behavior and function of circulatory system components. The mental models found by the experimenters are shown in Figure 39: the first

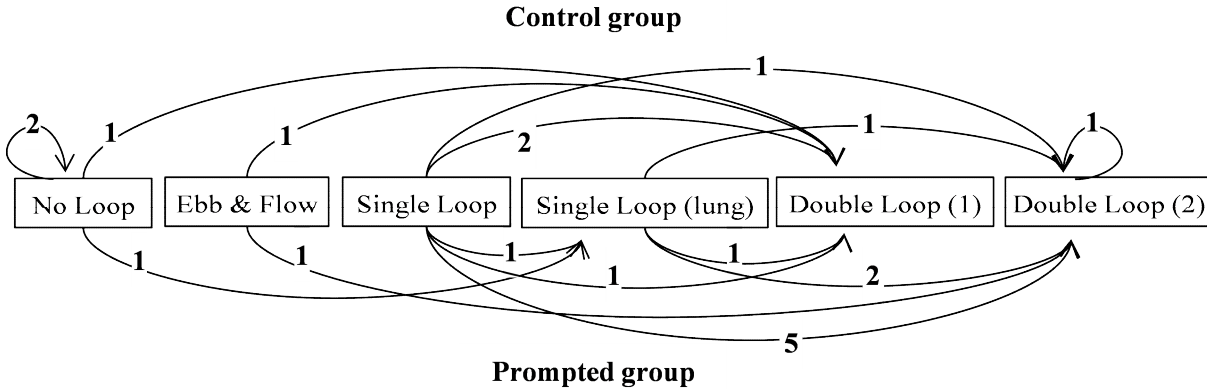
five are incorrect, and the final “double loop (2)” model is a correct but simplified model. We describe each from left to right:



**Figure 39: Student models of the human circulatory system from Chi et al. (1994a).**

1. *No loop*: blood flows from a single-chambered heart to the body and does not return.
2. *Ebb and flow*: blood flows from heart to the body and returns to the heart through the same blood vessels.
3. *Single loop*: blood flows from heart to body through one set of vessels and returns to the heart through an entirely different set of vessels.
4. *Single loop (lung)*: blood flows in a heart-lung-body or heart-body-lung cycle and the lungs play a role in oxygenating blood.
5. *Double loop (1)*: blood flows directly from heart to both lungs and body, and blood returns directly to the heart from the lungs and body.
6. *Double loop (2)*: same as double loop (1), except the heart has four chambers, blood flows top-to-bottom through the heart, and at least three of the following:
  - Blood flows from right ventricle to lungs
  - Blood flows from lungs to left atrium
  - Blood flows from left ventricle to body
  - Blood flows from body to right atrium





**Figure 40: Transitions between pretest and posttest models for control and prompted groups in Chi et al. (1994a). Numbers indicate the number of students who made the given transition. See Figure 39 for an illustration of each mental model.**

The experimenters found that the prompted group experienced a significant gain in learning relative to the control group, and that prompted students who self-explained most frequently achieved the “double loop (2)” model on the posttest. In total, 33% of the control group and 66% of the prompted group reached the correct mental model at the posttest. Results are summarized in Figure 40, with respect to the models shown in Figure 39.

Figure 40 shows that some students in the control group who started with the same model on the pretest ended with different models in the posttest. This is indicated by the fork at “No Loop” (i.e., two of these students end at “No Loop,” and the remaining student transitions to “Double Loop (1)”), and the fork at “Single Loop” (i.e., two of these students transition to “Double Loop (1)” and the remaining student transitions to “Double Loop (2)”). This means that factors other than the starting model affect students’ learning on this task. We broadly refer to these factors as *individual differences*. Students in the control group were largely left to learn according to their individual learning strategies, while students in the prompted group were influenced by prompting of the experimenter. Our simulation attempts to capture (1) the

individual differences of the control group using different explanation evaluation strategies and (2) the majority of the prompted group using a single explanation evaluation strategy.

## 7.2 Simulating the self-explanation effect

This simulation is laid out similarly to the previous chapter's simulation. The input to the system includes: (1) starting domain knowledge; (2) a single preference ranking<sup>37</sup> for computing preferences over explanations; and (3) a sequence of scenario microtheories containing information from a textbook passage. The information in the passage was hand-translated by me into predicate calculus. Items 1 and 2 vary across simulation trials to simulate different students, and item 3 is constant over all trials. Each trial of this simulation proceeded as follows:

1. Begin the trial with domain knowledge specific to one of the six mental models shown in Figure 39. No explanations are present.
2. Construct explanations for all blood flows believed to exist in the domain theory.
3. Generate an influence graph of all flows of blood, oxygen, and carbon dioxide from the union of preferred explanations. This validates the initial circulatory model, for comparison with student pretests.
4. Incrementally integrate textbook information about the circulatory system from a sequence of scenario microtheories.
5. After all of the textbook information has been integrated, generate influence graphs again from the union of preferred explanations as done in step (3). This determines the final circulatory model, for comparison with student posttests.

---

<sup>37</sup> See section 4.6.1 for how preference rankings affect explanation preferences.

```

ModelFragment ContainedFluid
Participants:
  ?con Container (containerOf)
  ?sub StuffType (substanceOf)
Constraints:
  (physicallyContains ?con ?sub)
Conditions:
  (greaterThan (Amount ?sub ?con) Zero)
Consequences:
  (qprop- (Pressure ?self) (Volume ?con))

ModelFragment FluidFlow
Participants:
  ?source-con Container (outOf-Container)
  ?sink-con Container (into-Container)
  ?source ContainedFluid (fromLocation)
  ?sink ContainedFluid (toLocation)
  ?path Path-Generic (along-Path)
  ?sub StuffType (substanceOf)
Constraints:
  (substanceOf ?source ?sub)
  (substanceOf ?sink ?sub)
  (containerOf ?source ?source-con)
  (containerOf ?sink ?sink-con)
  (permitsFlow ?path ?sub
    ?source-con ?sink-con)
Conditions:
  (unobstructedPath ?path)
  (greaterThan (Pressure ?source)
    (Pressure ?sink))
Consequences:
  (greaterThan (Rate ?self) Zero)
  (i- (Volume ?source) (Rate ?self))
  (i+ (Volume ?sink) (Rate ?self))

```

When a container *con* physically contains a type of substance *sub*, a contained fluid exists. When there is a positive amount of *sub* in *con*, the volume of *con* negatively influences the pressure of this contained fluid.

When two contained fluids – a *source* and a *sink* – are connected by a *path*, and both are of the same type of substance, a fluid flow exists. When the *path* is unobstructed and the pressure of *source* is greater than the pressure of *sink*, the rate of the flow is positive and it decreases the volume of *source* and increases the volume of *sink*.

**Figure 41: ContainedFluid (above) and FluidFlow (below) model fragments used in the simulation. English interpretations of each model fragment (at right).**

We use the simulation's influence graphs from steps (3) and (5) to assess the simulation's learning and compare it to the mental model transformations of Chi et al.'s students in Figure 40. We have already described the explanation construction procedures in detail: section 4.4 describes how an explanation of heart-to-body blood flow is constructed in this simulation. This is the essence of simulation step (2) above. Additionally, Chapter 6 describes how influence graphs are constructed from multiple preferred explanations, which is the essence of steps (3) and (5) in this simulation. We therefore concentrate on step (4) of the simulation: incrementally integrating textbook information.

### 7.2.1 Explanandums: situations that require an explanation

Unlike the simulation in the last chapter, the explanandums in this simulation are not single statements. Rather, each explanandum describes a single situation, such as:

```
(isa naiveH2B PhysicalTransfer)

(outOf-Container naiveH2B heart)

(into-Container naiveH2B body)

(substanceOf naiveH2B Blood)
```

These four statements describe a situation called `naiveH2B`. The `isa` statement identifies it as a `PhysicalTransfer` instance, and the `outOf-Container`, `into-Container`, and `substanceOf` statements identify the entities that fill these roles of `naiveH2B`. Although the situation is described across four statements, the situation itself (`naiveH2B`) is the explanandum. Consider another explanandum situation called `leftH2B`:

```
(isa leftH2B PhysicalTransfer)

(outOf-Container leftH2B l-heart)

(into-Container leftH2B body)

(substanceOf leftH2B Blood)
```

Using multiple statements to describe explanandums allows us to describe events with incomplete information. A more complete account of flow would also mention the paths through which the substance travels from source to destination. The path is, after all, a component of our

FluidFlow model fragment in Figure 41. Formal instruction does not always provide all of the information about the components of a natural system, especially when systems are described from the top-down. For example, consider the following sentence from the textbook passage used by Chi et al.:

“Blood returning to the heart [from the body]... enters the right atrium.”

A more complete passage would mention the superior and inferior vena cava, but these are omitted, perhaps to keep focus on the more general function and structure of the system. Consequently, students must assume the existence of a flow path from the body to the right atrium. Likewise, our simulation assumes the existence of entities to fill the roles of model fragments when necessary, using the abductive mechanism described in section 4.4.

### **7.2.2 Constructing explanations to generate the pre-instructional model**

When a simulation trial begins, there are no justifications or explanations in the system. The simulation has the following information in its domain knowledge microtheory: (1) a set of model fragments including those in Figure 41; (2) propositional beliefs about the structure of the circulatory system; and (3) a set of explanandum situations (described above) pertaining to a single model of the circulatory system shown in Figure 39. For example, a simulation trial that begins with the “single loop” model contains the following explanandum situations:

- Blood flows from the heart to the body (i.e., *naiveH2B*).
- Blood flows from the body to the heart.



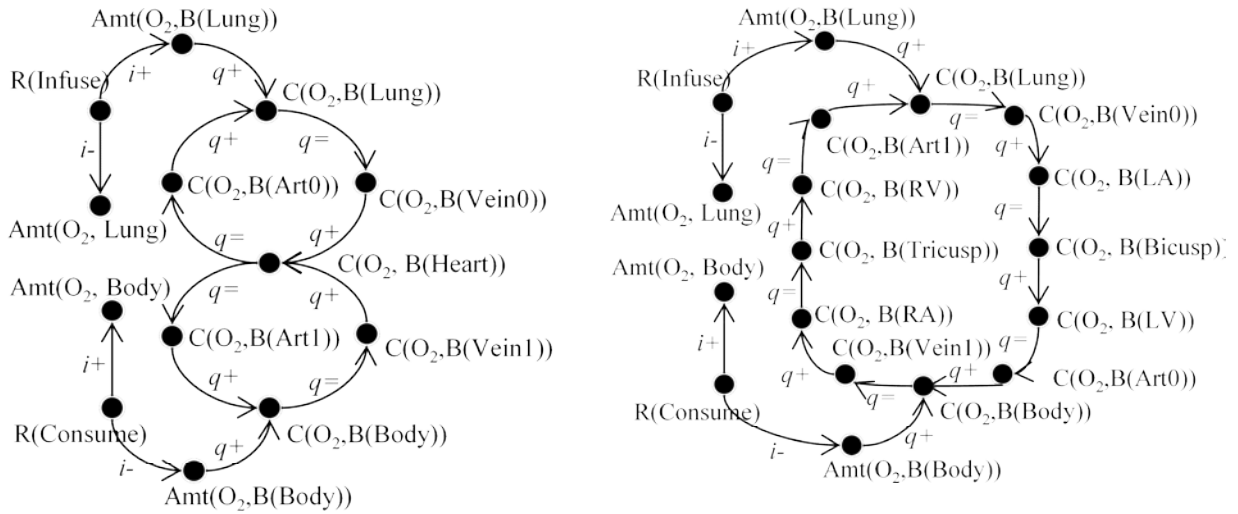
information about paths. The existence of a path will be assumed (i.e., without committing to a specific blood vessel or pathway) for the flow.

All of the starting explanandums and propositional beliefs are contextualized within scenario microtheories.<sup>38</sup> Each of these scenario microtheories is tagged as a starting microtheory (i.e., it existed prior to instruction) by labeling the `informationSource` of the microtheory as `Self-Token` (i.e., the symbol denoting the simulation itself). This is important, since the simulation will later resolve explanation competition based on the `informationSource` of the constituent beliefs.

The next step is to construct an explanation for each starting explanandum. The system automatically detects explanandums by querying for situations that match a specific pattern: descriptions of processes (e.g., blood flow, oxygen consumption) that are not themselves model fragment instances. For each explanandum, the system uses the *justify-explanandum* procedure and subsequent *justify-process* procedure to construct an explanation, both of which are described in Chapter 4. Consider the simple case of starting with the “no loop” student model. Figure 42(a) shows the system’s state prior to explaining `naiveH2B`, and Figure 42(b) shows the same portion of the network after explanation  $x_0$  is constructed for `naiveH2B`.

---

<sup>38</sup> See section 4.2 for discussion of scenario microtheories.



**Figure 43: Influence graphs generated by the system to describe the relative concentrations, infusion, and consumption of Oxygen. Left: using “Double loop (1)” model. Right: using “Double loop (2)” model.**

**Key:  $R(x)$ =Rate of process of type  $x$ ;  $\text{Amt}(x, y)$ =Amount of  $x$  in  $y$ ;  $C(x, y)$ =Concentration of  $x$  in  $y$ ;  $B(x)$ =Blood in region  $x$ ;  $(R/L)A$ =R-/L-Atrium;  $(R/L)V$ =R-/L-Ventricle.**

### 7.2.3 Determining the simulation’s circulatory model

Students in Chi et al. were asked to draw the blood flow in the human circulatory system as part of their pretest and posttest assessment. We assess our simulation’s circulatory model twice: (1) after explaining the starting explanandums and (2) after integrating the textbook information. Both of these assessments are conducted by having the system automatically generate influence graphs. This is accomplished with the following steps:

1. Find all explanandums  $M$  in the adopted domain knowledge microtheory  $\mathbb{D}_a$  that describe the transfer, consumption, or infusion of blood, Oxygen, or Carbon Dioxide.
2. Using the explanandum mapping  $\mathbb{E}$  described in Chapter 4, locate identify the preferred explanations  $X$  for each explanandum  $M$ .



3. Using all of the beliefs of explanations  $X$ , construct an influence graph describing transfers, consumption, or infusion of blood, Oxygen, or Carbon Dioxide.

This produces between one and three influence graphs, since all student circulatory models in Chi et al. describe the transfer of blood, but not all of them describe oxygen and carbon dioxide (see Figure 39 and the corresponding coding criteria). Influence graphs describing Oxygen are shown in Figure 43 for two different circulatory models: “double loop (1)” (left) and “double loop (2)” (right). The “double loop (1)” graph describes oxygenated blood flowing from the lung to the heart via a vein pathway  $Vein0$ , where it mixes with deoxygenated blood from the body, flowing to the heart via vein pathway  $Vein1$ . The “double loop (2)” graph has no such mixture.

Influence graphs constitute a partial comparison to the students in Chi et al., since the students also completed a quiz about the function of the circulatory system. Influence graphs effectively map the simulation’s circulatory model onto the space of student models in Figure 39, but it does not directly measure the simulation’s knowledge about the function of the circulatory system and its impact on human nutrition.

#### **7.2.4 Integrating textbook information**

At this point, the system has (1) constructed explanations for each starting explanandum and (2) generated an influence graph to describe its initial circulatory model. This section describes how textbook information is integrated to incrementally transform this circulatory model. The portion of the textbook passage used by our simulation is listed in the appendix. For the

remainder of this section, we suppose that the simulation started with the “no loop” model of the circulatory system discussed above.

We present the textbook information in small increments, as a sequence of scenario microtheories. Unlike the starting scenario microtheories with `Self-Token` as the source of information, these microtheories are encoded with source `Instruction`. Otherwise, they only vary in content. The first sentence from the textbook passage describes the general structure of the heart: “The septum divides the heart lengthwise into two sides.” The corresponding scenario microtheory contains the following facts:

```
(isa septum Septum)
(physicallyContains heart septum)
(isa l-heart (LeftRegionFn Heart))
(isa r-heart (RightRegionFn Heart))
(partitionedInto heart l-heart)
(partitionedInto heart r-heart)
(between l-heart r-heart septum)
(rightOf r-heart l-heart)
```

First, the adopted domain knowledge microtheory  $\mathbb{D}_a$  is added as a child of the new scenario microtheory so that the new information is visible from this context. This scenario microtheory does not contain an explanandum, so nothing new requires an explanation. However, new entities are described, including `septum`, `l-heart`, and `r-heart`. These entities did not exist in the simulation’s “no loop” circulatory system model. Consequently, the simulation uses the

preference rules described in Chapter 4 to encode preferences over entities, where possible. The following preferences are computed:

1.  $(\text{isa heart Heart}) <_c^s (\text{isa l-heart (LeftRegionFn Heart)})$
2.  $(\text{isa heart Heart}) <_c^s (\text{isa r-heart (RightRegionFn Heart)})$
3.  $(\text{isa heart Heart}) <_c^i (\text{isa l-heart (LeftRegionFn Heart)})$
4.  $(\text{isa heart Heart}) <_c^i (\text{isa r-heart (RightRegionFn Heart)})$
5.  $(\text{isa l-heart (LeftRegionFn Heart)}) <_c^n (\text{isa heart Heart})$
6.  $(\text{isa r-heart (RightRegionFn Heart)}) <_c^n (\text{isa heart Heart})$

Preferences 1 and 2 are specificity (*s*) preferences, and are computed based on the specificity, since heart is partitioned into the subregions r-heart and l-heart. Preferences 3 and 4 are instruction (*i*) preferences: since l-heart and r-heart are both comparable to heart for specificity and are supported by instruction (i.e., with information source Instruction), they are preferred in this (*i*) dimension. Finally, preferences 5 and 6 are prior knowledge (*n*) preferences: since l-heart and r-heart are both comparable to heart for specificity, but neither were present prior to instruction (as was heart, with information source Self-Token), heart is preferred in this (*n*) dimension.

The next scenario microtheory describes the sentence “The right side pumps blood to the lungs, and the left side pumps blood to other parts of the body,” and contains the following statements:

$(\text{physicallyContains r-heart Blood})$

```
(physicallyContains l-heart Blood)
```

```
(physicallyContains lung Blood)
```

```
(isa rightH2L PhysicalTransfer)
```

```
(outOf-Container rightH2L r-heart)
```

```
(into-Container rightH2L lung)
```

```
(substanceOf rightH2L Blood)
```

```
(isa leftH2B PhysicalTransfer)
```

```
(outOf-Container leftH2B l-heart)
```

```
(into-Container leftH2B body)
```

```
(substanceOf leftH2B Blood)
```

This scenario microtheory describes two processes: `rightH2B` describes blood flow from right-heart to lungs and `leftH2B` describes blood flow from left-heart to body. Preferences can be computed between explanandums provided the following rule:

*If one explanandum  $e_1$  has one or more role fillers (e.g., `l-heart` in `leftH2B`) that are preferred for specificity  $<_c^s$  over the corresponding role filler of another explanandum  $e_2$  (e.g., `heart of naiveH2B`), and all other corresponding role fillers that are not preferred are identical, encode a specificity preference  $e_1 <_c^s e_2$ .*

This rule is domain general, since it describes specificity over all events; not just physical transfers and blood flows. Recall that in our example, the simulation starts with the “no loop”

model. This means that before encountering `leftH2B`, the network contains only explanandum `naiveH2B`. This means that the following preference will be computed:

1.  $\text{naiveH2B} <_c^s \text{leftH2B}$
2.  $\text{naiveH2B} <_c^i \text{leftH2B}$
3.  $\text{leftH2B} <_c^n \text{naiveH2B}$

These indicate that (1) `leftH2B` is more specific than `naiveH2B`, (2) `leftH2B` is supported by instruction and `naiveH2B` is not, and (3) `naiveH2B` was present prior to reading, and `leftH2B` was not. The simulation next automatically constructs and evaluates explanations for new explanandums `leftH2B` and `rightH2L`. We describe how `leftH2B` is explained.

Since our discussion focuses on a simulation trial with the “no loop” model, the network contains only an explanation for `naiveH2B`, as in Figure 44(a). To explain `leftH2B`, the simulation invokes *justify-explanandum* using `leftH2B` as the explanandum argument. This constructs an explanation for `leftH2B` using knowledge about `l-heart` from the first scenario microtheory. This explanation  $x_1$  is shown in Figure 44(b), coexisting with the explanation  $x_0$  for `naiveH2B`. Notice that in Figure 44(b), some of the preferences computed above are shown. Moreover, since the explanandum `leftH2B` is preferred for specificity over `naiveH2B`, any explanation for `leftH2B` (e.g., new explanation  $x_1$ ) also explains `naiveH2B`. This is reflected in Figure 44(b).

According to our discussion of explanation competition in Chapter 4, any two explanations that explain the same explanandum(s) are in competition. In Figure 44(b),  $x_0$  and  $x_1$  both explain `naiveH2B`, so rule-based preferences are used to compute preferences between  $x_0$  and  $x_1$ . These

preferences and a preference aggregation function will determine which explanation will be assigned to `naiveH2B` in the explanandum mapping  $\mathbb{E}$ . The following preferences are computed as follows, using the above domain-level preferences already discussed above. Let  $c_{heart}$  be the `ContainedFluid` instance with participants  $\langle ?sub, blood \rangle$  and  $\langle ?con, heart \rangle$ , and let  $c_{left}$  be the `ContainedFluid` instance with participants  $\langle ?sub, blood \rangle$  and  $\langle ?con, l-heart \rangle$ . Similarly, let  $f_{heart}$  be the `FluidFlow` instance with binding  $\langle ?source-con, heart \rangle$ , and let  $f_{left}$  be the `FluidFlow` instance with binding  $\langle ?source-con, l-heart \rangle$ .

1.  $c_{heart} <_{mfi}^i c_{left}$
2.  $c_{heart} <_{mfi}^s c_{left}$
3.  $c_{left} <_{mfi}^n c_{heart}$
4.  $f_{heart} <_{mfi}^i f_{left}$
5.  $f_{heart} <_{mfi}^s f_{left}$
6.  $f_{left} <_{mfi}^n f_{heart}$

These preferences over model fragment instances are used to compute three explanation-level preferences between the prior heart-to-body explanation  $x_0$  and the new left-heart-to-body explanation  $x_1$ :

1.  $x_0 <_{xp}^i x_1$
2.  $x_0 <_{xp}^s x_1$
3.  $x_1 <_{xp}^n x_0$

One of these explanations must be mapped to `naiveH2B` as its preferred explanation in the explanandum mapping  $\mathbb{E}$ ; however, the three explanation-level preferences above describe a preference cycle. Cycles are resolved using a preference aggregation function, as described in Chapter 4. The preference aggregation function is given a preference ranking which is an ordering over preference dimensions  $\{s, i, n, r\}$ , where a dimension earlier in the ordering is more important than a dimension later in the ordering. The aggregation function begins with the first dimension of the preference ranking and honors those preferences, and then honors each of the preferences in the next dimension as long as it does not create a cycle, and so-on for all dimensions. If  $n$  precedes  $s$  and  $i$  in the preference ranking, the system will prefer  $x_0$  over  $x_1$ ; otherwise,  $x_1$  will be preferred. This ultimately determines which explanation will be mapped to `naiveH2B` in  $\mathbb{E}$ , and thereby affects how the system will explain blood flow on the posttest.

The preference ranking also applies to explanandums: if  $n$  precedes  $s$  and  $i$  in the preference ranking, then (based on the above preferences) the explanandum `naiveH2B` will be preferred over `leftH2B`. Explanandum preferences are used for pruning – if explanandum  $a$  is preferred over explanandum  $b$  then explanandum  $b$  is not used for problem solving, question answering, or generating an influence graph.

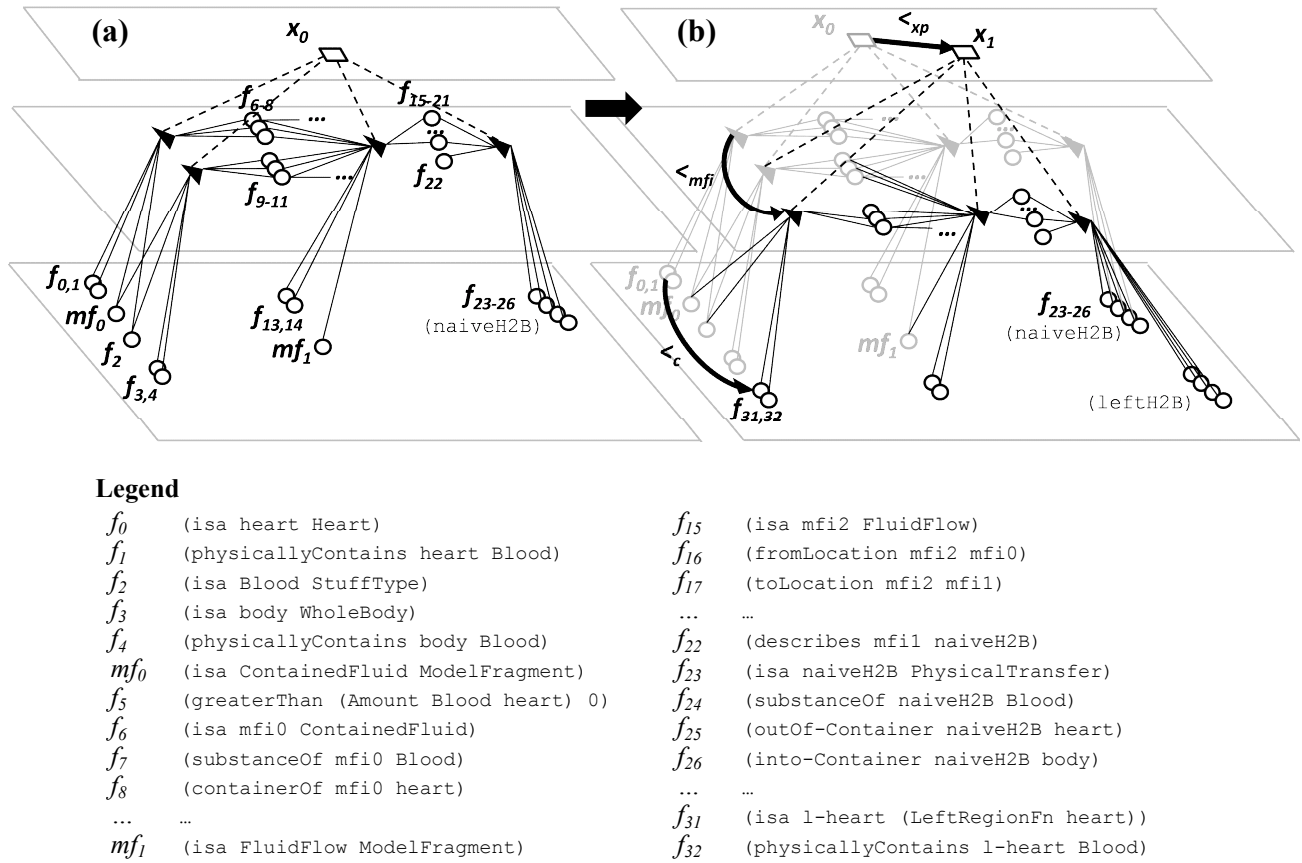
Since preferences are computed over entities and model fragment instances, the preference ranking ultimately affects the granularity and terminology of the explanation. For example, if the prior knowledge preference  $n$  is first, the system will prefer pre-instructional entities (e.g., heart) over more specific entities and regions thereof (e.g., left-heart, right-heart, left-ventricle, right-ventricle, left-atrium, and right-atrium), and this will be reflected in the choice of explanations. This is an example of how we can model resistance to change: favoring pre-

instructional entities over new entities whenever a choice is available makes the system selectively incorporate new information into its qualitative models. Conversely, if instruction (*i*) is first in the preference ranking, then textbook information will displace pre-instructional information in preferred explanations.

As mentioned above, the preference ranking is an input to the simulation, so each trial has a single preference ranking that it uses throughout learning. By varying this preference ranking, we can change the outcome of learning and thereby simulate different students, including individual differences. In this simulation the preference ranking is an approximation of a student's learning strategy. Recall that some students in the control group who started with the same mental model in the pretest diverged in their mental models at the posttest. As we will show, the preference ranking helps account for these differences.

We have described how our computational model integrates new information by constructing explanations and computing preferences. The content of each scenario microtheory varies, but the explanation construction and evaluation mechanisms are constant.





**Figure 44: Portion of explanation-based network. (a): After explaining blood flow from heart to body (*naiveH2B*). (b): After explaining blood flow from the left-heart to the body (*leftH2B*), with preferences across concepts ( $<_c$ ), model fragment instances ( $<_{mfi}$ ), and explanations ( $<_{xp}$ ).**

### 7.2.5 Assuming model participants

In some cases, an explanandum is presented to the system when the system does not have complete information. Consider the sentence “Blood returning to the heart [from the body], which has a high concentration of carbon dioxide and a low concentration of oxygen, enters the right atrium.” The corresponding scenario microtheory contains the following statements:

```
(isa bloodToAtrium-Right FlowingFluid)

(substanceOf bloodToAtrium-Right Blood)

(outOf-Container bloodToAtrium-Right body)
```

```

(into-Container bloodToAtrium-Right right-atrium)

(valueOf ((ConcentrationOfFn Oxygen) bloodToAtrium-Right)
  (LowAmountFn (ConcentrationOfFn Oxygen)))

(valueOf ((ConcentrationOfFn CarbonDioxide) bloodToAtrium-Right)
  (HighAmountFn (ConcentrationOfFn CarbonDioxide)))

```

From this description of the blood that flows from the body to the right atrium, the system can gather most of the participants of a `FluidFlow`: the substance is blood; the source container is the body; the destination container is the right atrium; and the `ContainedFluid` instances corresponding to these containers are the source and destination fluids. However, no entity is included in this scenario microtheory that conforms to the collection and constraints of the `?path` participant of this `FluidFlow`, and the agent may not know of any entity that permits blood flow from the body to the right atrium.

As discussed in section 4.4, the model formulation algorithm assumes the existence of entities to fill these participant slots. When explaining the situation `bloodToAtrium-Right`, suppose the model formulation algorithm cannot bind a known entity to the `?path` participant slot which corresponds to the role `along-Path` of `FluidFlow` (see Figure 41 for details). The algorithm still creates a new `FluidFlow` model fragment instance with a unique symbol such as `mfi5`, and will construct an entity with a skolem term (discussed in Chapter 5) such as `(SkolemParticipant mfi5 along-Path)`. This indicates that this entity was assumed as a participant of `mfi5` for the role `along-Path`. The following two assertions are inferred as well:

```

(isa (SkolemParticipant mfi2 along-Path) Path-Generic)

(permitsFlow (SkolemParticipant mfi5 along-Path) Blood body right-atrium)

```

These statements describing the assumed entity will be part of the resulting explanation. This allows the system to construct qualitative models with partial information.

### 7.3 Simulation results

Here we describe the results of our simulation. Each trial of our simulation varied across three parameters: (1) the system's starting model, one of the six shown in Figure 39; (2) whether or not the system constructs explanations for new explanandums, and (3) the preference ranking. Varying the latter two settings makes two psychological assumptions which we discuss in the next section.

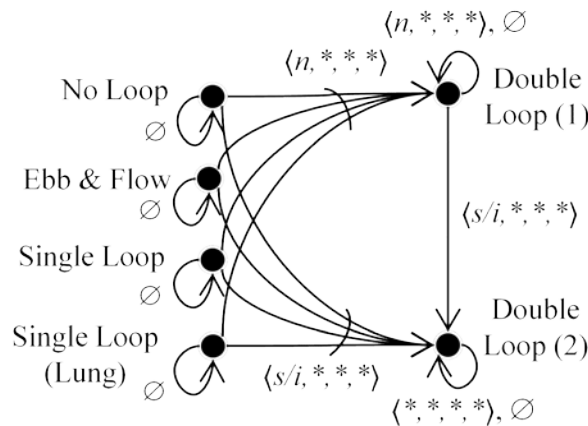
Each trial proceeds in the same fashion: (1) validate the starting (pretest) model with influence graphs; (2) incorporate the textbook information via a sequence of scenario microtheories as described above; and (3) determine the ending (posttest) model with influence graphs.

The results are shown in Figure 45. Each node in the figure corresponds to a student circulatory model in Figure 39, and each labeled arrow between circulatory models indicates that the simulation transitions from one model to another using the labeled preference ranking. For instance, by engaging in full self-explanation with preference ranking  $\langle s/i, *, *, * \rangle$  (i.e., the last three preferences are irrelevant, provided the first is either instruction or specificity), the simulation could transition to the correct “double loop (2)” circulatory model from any initial model. Further, recall that using ranking  $\langle n, *, *, * \rangle$  biases the system to favor explanations that use prior (i.e., starting model) entities, such as `heart`, over comparable entities encountered via instruction, such as `left-ventricle`. This resulted in the simulation learning the most

popular final model in Chi's control group, "double loop (1)" (Figure 43, left). This mode uses heart instead of the more specific regions of the heart used in "double loop (2)" (Figure 43, right). By disabling explanation construction (Figure 45,  $\emptyset$ ), the system always remained at its initial circulatory model.

Individual differences in the control group were modeled using different preference orderings  $\langle n, *, *, * \rangle$  (4 students),  $\emptyset$  (3 students), and  $\langle s/i, *, *, * \rangle$  (2 students). The prompted students were modeled using preference ordering  $\langle s/i, *, *, * \rangle$  (8 students) and  $\langle n, *, *, * \rangle$  (2 students). The remaining two prompted students were not modeled by the system. Both transitioned to the "single loop (lung)" model – one from "no flow" and one from "single loop." The inability of our system to generate these transitions may be due to representation differences, either in the starting knowledge or in the representation of the instructional passage. We discuss this further in the next section.

By varying the initial circulatory model, the preference rankings, and whether or not the system constructs explanations, the system was able to capture 19 out of 21 (>90%) of student model transitions in the psychological data. Individual differences in the control group were



**Figure 45: Circulatory model transitions for all simulation trials.**

captured by three parameter settings, and the majority of the prompted group was modeled by encoding a preference for explanations that contained specific and instructional concepts,  $\langle s/i, *, *, * \rangle$ .

## 7.4 Discussion

We have simulated self-explanation using model formulation, metareasoning, and epistemic preferences. By altering its preference rankings, we are able to affect how the system prioritizes its knowledge and integrates new information.

Our simulation trials vary with respect to (1) whether the system explains textbook information, and (2) the preference ranking it uses to evaluate explanations. Since our model learns by explaining, changing setting (1) to disable explanation construction prohibits learning. This means that some simulation trials will not integrate any textbook information, which therefore assumes that some students do not learn from reading the textbook passage. This was indeed the case for students in Chi et al.’s control group, since two students in Figure 40 started and ended with the same incorrect model.

Varying the preference ranking assumes that students have different strategies for assimilating information from text. This must be the case, because we cannot explain the learning patterns of the control group in Figure 40 based on their starting model alone: of the three students in the control group who began with the “single loop” model, two of them transitioned to “double loop (1),” and one transitioned to “double loop (2).” Consequently, the system must capture these individual differences with at least two different learning strategies, which we model using preference rankings.

For autonomous learning systems and for modeling human learning over multiple reading tasks, the preference ranking might need to be more dynamic, reflecting depth of experience versus the credibility of the source. Nevertheless, the simulation demonstrates good coverage of the psychological data.

We have shown that the space of preference rankings  $\langle s/i, *, *, * \rangle$  results in the correct model from any initial model for this task. This may not be the case for other domains and for students whose mental models are flawed by containing extraneous components. For instance, in this study, textbook entities (e.g., `left-ventricle`) were generally at least as specific as the entities in students' initial models (e.g., `heart`). This means that the partial orderings  $<_c^s$  and  $<_c^i$  had a near-perfect correspondence over entities.<sup>39</sup> We can imagine other cases where this might not be true. For example, a student may have erroneous initial beliefs about a `left-ventricle-basin` region of the `left-ventricle`. Since this region does not actually exist, the initial, incorrect entity is more specific than the instructional entity. Any preference ranking that places specificity before instruction, such as  $\langle s, *, *, * \rangle$ , would retain the `left-ventricle-basin` misconception in the posttest. The opposite would be true if instruction is ranked over specificity.

This simulation supports our hypothesis that constructing and evaluating explanations can model the benefits of self-explanation. Additionally, the knowledge representation was sufficient to explain the flows of blood, CO<sub>2</sub>, and O<sub>2</sub> in the pretests and posttests in ways that are compatible with students' explanations, so that the system's qualitative models are comparable

---

<sup>39</sup> Specificity and instruction do not overlap perfectly in this study. Consider a student who already knows about the left atrium and left ventricle (the two sub-regions of the left heart): when they read about the left heart early in the text, the entities in their initial mental model are temporarily more specific than the entities in the textbook model.

to students' mental models. This provides evidence for our claim that compositional qualitative models can simulate human mental models.

While our methods were sufficient to simulate the majority of the students, two of the students in the self-explanation group were not captured. These students both used the “single loop (lung)” model at the posttest – one transitioned there from the “no flow” model and the other from the “single loop” model. This suggests that our model of self-explanation is incomplete. These students might have hypothesized system components based on the function of the system. If informed that (1) the lungs oxygenate the blood and that (2) the purpose of the circulatory system is to provide the body with oxygen and nutrients, one might infer that blood flows directly from the lungs to the body.

In our simulation, self-explanation generates network structure and preferences, which makes new knowledge available for later problem-solving. When we disabled self-explanation (Figure 45,  $\emptyset$ ), the new knowledge was unavailable for later use.

We have shown how existing models are recombined to explain new situations and accommodate new information. This has simulated how people revise and reason with mental models: new domain elements are acquired through simulated instruction, and conceptual change is achieved by combining elements of domain knowledge into new, preferred models. This does not account for the revision of categories and model fragments themselves. We simulate this type of conceptual change in the next chapter.

## Chapter 8: Revising a category of force when explanations fail

Naïve theories of force are some of the most widely-studied misconceptions, and are also some of the most resilient to change. The questions of how intuitive theories of force are learned, represented, and revised are debated in the literature, but there is some agreement that they are mechanism-based (McCloskey, 1983; Ioannides & Vosniadou, 2002; diSessa et al., 2004) and learned and reinforced by experience (Smith, diSessa, & Roschelle, 1994).

Here we describe a simulation that creates and revises a force-like category to explain a sequence of observations.<sup>40</sup> Categories and model fragments are created and revised upon explanation failure. After each observation, the system completes a questionnaire from previous psychology experiments (Ioannides & Vosniadou, 2002; diSessa et al., 2004) so we can compare its answers to those of students. We then plot the system's learning trajectory against student data to show that the simulation can learn and transition between student-like categories of force. The system transitions between mutually inconsistent specifications of a force-like category, along a humanlike trajectory. This simulation thereby provides evidence for claims 1 and 3 of this dissertation:

***Claim 1:*** Compositional qualitative models provide a consistent computational account of human mental models.

***Claim 3:*** Human mental model transformation and category revision can both be modeled by iteratively (1) constructing explanations and (2) using meta-level reasoning to select among competing explanations and revise domain knowledge.

---

<sup>40</sup> This builds upon the simulation described in Friedman and Forbus (2010).



Like the simulations in Chapters 6 and 7, this simulation constructs and evaluates explanations to simulate human conceptual change. However, this simulation also uses heuristics to revise its model fragments and categories when it fails to explain an explanandum, and then it attempts explanation again. This conforms to the following pattern of events:

1. A new explanandum within a scenario requires an explanation.
2. No explanation can be constructed that is consistent with the scenario. We call this an *explanation failure*.
3. The system finds heuristics that are applicable to the present failure mode.
4. Applicable heuristics are sorted by their estimated complexity of change to domain knowledge.
5. Beginning with the heuristic that incurs the least change, execute the heuristic to add or revise domain knowledge as necessary. If explanation still fails, repeat with the next heuristic.

After each explanandum within a scenario is explained, MAC/FAC is used to retrieve a similar, previously explained scenario. If the two scenarios are sufficiently similar, discrepancies are detected between the new and previous scenario, and are explained using the same process as above, using heuristics to revise knowledge as necessary. We describe both of these explanation-driven processes of change in detail below. First, we outline the results of Ioannides & Vosniadou (2002) and diSessa et al. (2004), which serve as the bases for comparison in this simulation.

## 8.1 Assessing the changing meaning of force in students

Ioannides & Vosniadou (2002) conducted an experiment to assess students' ideas of force. They used a questionnaire of sketched vignettes which asked the student about the existence of forces on stationary bodies, bodies being pushed by humans, and bodies in stable and unstable positions. They concluded that several meanings of force were held by the students:

1. **Internal Force** (11 students): A force exists inside all objects, affected by size/weight.
2. **Internal Force Affected by Movement** (4 students): Same as **Internal Force**, but position/movement also affects the amount of force.
3. **Internal & Acquired** (24 students): A force exists due to size/weight, but objects acquire additional force when set into motion.
4. **Acquired** (18 students): Force is a property of objects that are in motion. There is no force on stationary objects.
5. **Acquired & Push-Pull** (15 students): Same as (4), but a force exists on an object, regardless of movement, when an agent pushes or pulls it.
6. **Push-Pull** (1 student): A force only exists when objects are pushed or pulled by an agent.
7. **Gravity & Other** (20 students): Forces of gravity, of push/pull, and acquired force when objects are moving.

8. **Mixed** (12 students): Responses were internally inconsistent, and did not fall within the other categories.

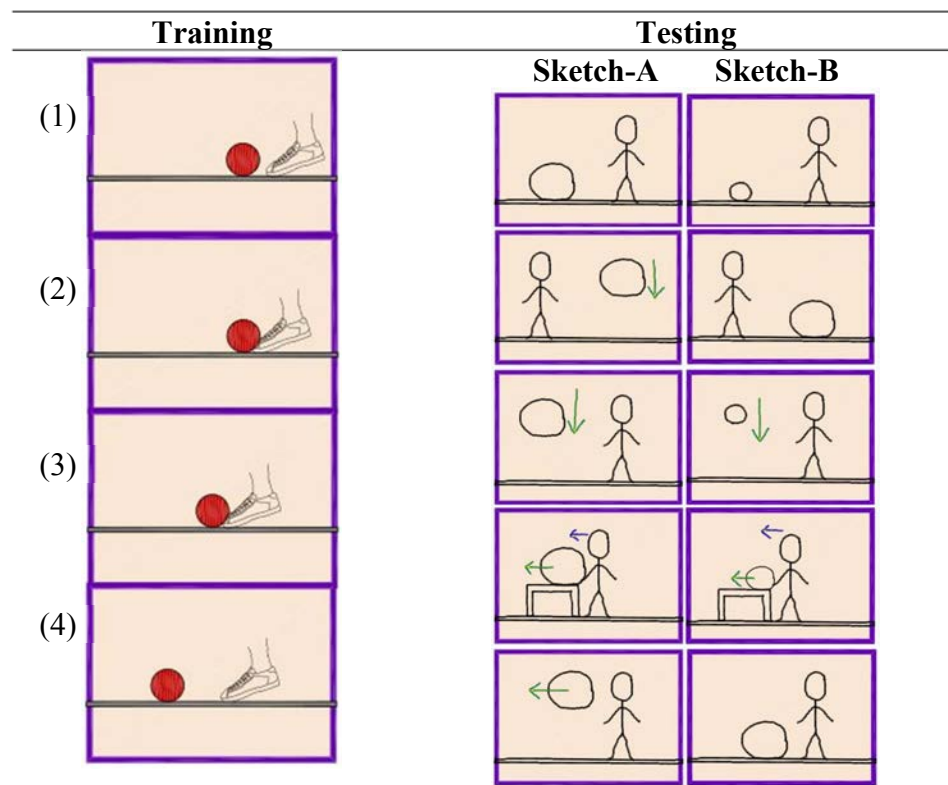
Meaning of Force	K	4th	6th	9th	Total
Internal	7	4			11
Internal/Movement	2	2			4
Internal/Acquired	4	10	9	1	24
Acquired		5	11	2	18
Acquired/Push-Pull			5	10	15
Push-Pull				1	1
Gravity/Other		3	1	16	20
Mixed	2	6	4		12

**Figure 46: Occurrences of meaning of force, by grade.**

The frequencies of responses by grade are listed in Figure 46. Though these data were gathered on different students across grades, they illustrate a trend: Kindergarteners favor the “Internal” meaning of force, and then transition through the “Internal & Acquired” meaning to the “Acquired” meaning. By grade 9, students tend to adopt the “Acquired & Push-Pull” and “Gravity & Other” meanings.

diSessa et al. (2004) conducted a replication of Ioannides & Vosniadou (2002) using a modified questionnaire, but was not able to reliably classify students using the same coding criteria. diSessa et al.’s conclusions include: (1) students do not form and transition between coherent theories (cf. Ioannides & Vosniadou, 2002); (2) rather, student theories are composed of small, contextualized, pieces of knowledge, some of which are idiosyncratic; and therefore (3) classifying each student into one of several coherent theories does not help us understand the processes by which students use and revise conceptual knowledge. diSessa et al.’s conclusions are consistent with the knowledge in pieces perspective discussed in Chapter 2.

Despite this controversy, the student data from Ioannides & Vosniadou’s study provides a clear basis for comparison for our simulation. Additionally, since our approach incorporates ideas from both knowledge in pieces (i.e., we represent domain knowledge with globally incoherent, composable elements) and theory theory (i.e., explanations are coherent aggregates of said elements), we have the opportunity to demonstrate how an agent with globally incoherent domain knowledge can transition through a trajectory of apparently coherent meanings of force.



**Figure 47: At left: a four-frame comic graph used as training data. At right: five of the ten questionnaire scenarios used as testing data.**

Ioannides & Vosniadou and diSessa et al. both used a sketch-based questionnaire to characterize each student’s concept of force. Ioannides & Vosniadou’s questionnaire varied slightly from diSessa et al.’s version, so we used the more recent and succinct (diSessa et al.) variation. The questionnaire contains ten scenarios, five of which are illustrated in Figure 47

(right). Each scenario contains two sketches (A and B) of a person and a rock, and the student is asked three questions:

1. What forces act on rock A?
2. What forces act on rock B?
3. Is the force on rock A same or different as the one on rock B?

One or more aspects vary between the A and B sketch within a scenario (e.g., the size of the rock, the size of the person, and the motion of the rock). This helps identify which variables determine the existence and magnitude of force, which ultimately determines the student's ideas of force.

### **8.1.1 Replicating the force questionnaire and approximating students' observations**

We sketched the questionnaire from diSessa et al. using CogSketch (illustrated in Figure 47, right). We use sketched annotations, as described in Chapter 5, to indicate pushing (blue arrows) and movement (green arrows) as indicated in the original questionnaire. We use the same coding strategy as diSessa et al. and Ioannides & Vosniadou to classify our simulation's meaning of force. To simplify coding, we interpret diSessa et al.'s question (3) as:

3. Which rock has greater force(s) acting on it, if they are comparable?

This allows us to query for an ordinal relationship (e.g., `greaterThan`, `lessThan`, or `equalTo`) between the quantities of force on the rocks.

We also use CogSketch to sketch comic graphs (see Figure 47, left), which are used as training data. These comic graphs are similar to those in Chapter 5, except they contain no annotations. Consequently, the system has to detect motion and infer force-like quantities independently. As mentioned above, the entire sketched questionnaire is interleaved after each comic graph training datum to determine which, if any, student meaning of force in Figure 46 is used by the system. Each simulation trial thus generates a sequence of force categories. We can plot this sequence of force categories against the student data in Figure 47 to determine whether the system's trajectory of learning follows a pattern within the results of Ioannides & Vosniadou (2002).

We next discuss how comic graphs are processed and explained by the simulation, and how heuristics are used to revise knowledge upon failure.

## **8.2 Learning by explaining new observations**

When the simulation is given a new comic graph as a training datum, it detects all quantity changes in the comic graph, such as movements along the x-axis. These quantity changes are explanandums, so the simulation must explain why each quantity change starts, persists, and stops. If no explanation can be constructed that is consistent within the scenario, then the system revises its domain knowledge until all quantity changes can be explained.<sup>41</sup> We discuss these operations in the order in which they occur, using the comic graph shown in Figure 47(left) to illustrate.

---

<sup>41</sup> When people encounter anomalies, they can ignore them altogether (Feltovich et al., 2001), hold them in abeyance, or exclude them from their domain theory (Chinn & Brewer, 1998). Our simulation's only response to anomaly is revision, so we expect rapid transition between concepts of force. We address this in the discussion section of this chapter.

<pre> Heuristic createDecreaseProcess Participants:   ?obj Entity   ?q Quantity Constraints:   (decreasing (?q ?obj)) Consequences:   (isa ?process ModelFragment)   (revise ?process (addParticipant ?e Entity))   (revise ?process (addConsequence (&gt; (Rate ?self) 0)))   (revise ?process (addConsequence (i- (?q ?e) (Rate ?self)))) </pre>	<pre> ModelFragment m<sub>1</sub> Participants:   ?e Entity Constraints:   nil Conditions:   nil Consequences:   (&gt; (Rate ?self) 0)   (i- (x-pos ?e) (Rate ?self)) </pre>
--	--

**Figure 48: Left: a heuristic `createDecreaseProcess` that automatically creates a new model fragment to explain a quantity decreasing. Right: Process model of leftward movement  $m_1$  automatically created with this heuristic.**

The simulation first finds quantity changes by comparing adjacent subsketches (e.g., subsketches 3 and 4 in Figure 47, left) using the spatial quantities encoded by CogSketch. If a quantity varies over a constant threshold (to account for unintentional jitter while sketching), a quantity change is encoded over that quantity for the transition. For example, in the 2→3 and 3→4 transitions, the x-coordinate of the ball decreases. Once the system computes all quantity changes within a comic graph, it must explain why each quantity change begins and ends.

For our discussion, suppose the simulation is explaining the ball’s movement as seen in the transitions 2→3→4. Suppose also that this is the first comic graph that the system has encountered. Since the simulation begins with no model fragments and no explanations, it will fail to explain the ball’s movement. Heuristics are used to revise and extend domain knowledge in order to accommodate this observation.

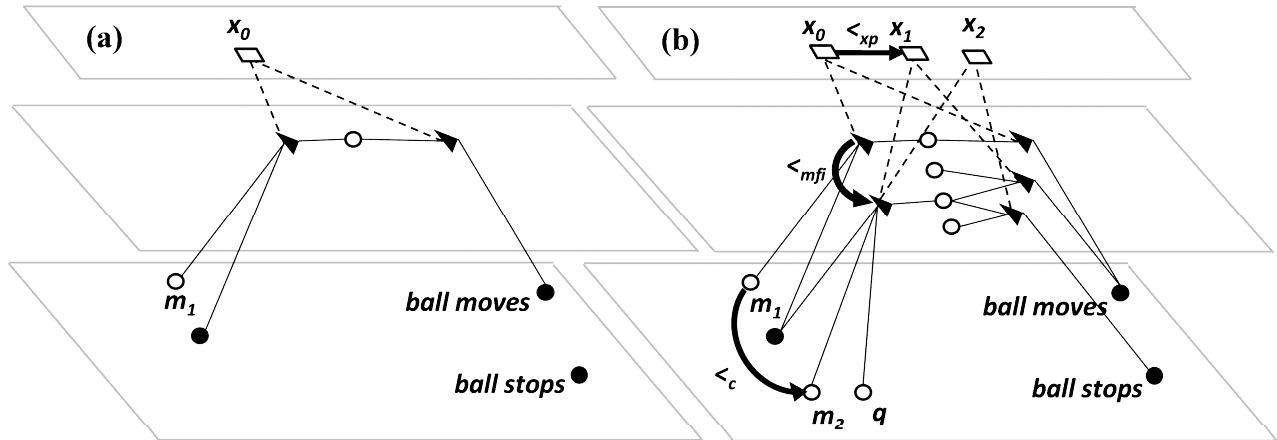
### 8.2.1 Declarative heuristics for failure-based revision

Like model fragments, heuristics are *declarative*. This means that the system can inspect them in order to decide which to use. To illustrate why this is important, suppose that the system is unable to explain an object’s motion, and two heuristics apply to the situation: (1) revise an

existing model fragment by adding a statement to its conditions; or (2) hypothesize a new, unobservable category that causes objects to resist motion, and then revise a model fragment to account for this. Which heuristic should the system choose? The psychology literature suggests that students make minimal changes to their theories when confronted with anomalous data (Chinn and Brewer, 1993), so our system makes the minimal change possible. It inspects heuristics to rate the amount of change they will incur, and sort them accordingly. Heuristics are defined using similar vocabulary as model fragments. Figure 48 (left) shows one such heuristic used by the system, which we will describe within our example.

Continuing our example in the previous section, suppose the simulation is given the comic graph of the foot kicking the ball to the left in Figure 47(left), and must explain the ball moving. Since the system begins without any model fragments or explanations, it fails to explain the ball's movement. It finds applicable heuristics by testing the participants and constraints of the heuristics. The heuristic `createDecreaseProcess` in Figure 48 (left) applies to this situation, since a quantity  $?q$  of an entity  $?obj$  is decreasing (i.e., a ball's x-axis position is decreasing). The consequences of this heuristic (1) add a new, empty model fragment  $?process$  to the domain knowledge microtheory and (2) revise  $?process$  so that it describes the corresponding quantity  $?q$  of an entity  $?e$  decreasing. This produces the process model  $m_l$  (Figure 48, right) which describes an object moving to the left. The ball's leftward movement can now be explained using  $m_l$ .





**Figure 50 (a) Model fragment  $m_1$  (Figure 48, right) explains the ball moving, but not the ball stopping. (b) After revising  $m_1$  as  $m_2$  (Figure 49, right),  $m_2$  explains both phenomena, and preferences are computed.**

The system now has a rudimentary model fragment  $m_1$  which describes objects – actually, all objects – moving continually to the left. The resulting network is shown in Figure 50(a). Next, the system must explain why the ball *stops* moving. Provided only model fragment  $m_1$ , this is not possible. The system must revise its knowledge again to resolve this next failure. This revision is illustrated in Figure 49, where the heuristic `addHiddenQtyCond` is used to revise  $m_1$  as  $m_2$ . The participants and constraints of heuristic `addHiddenQtyCond` assert that it

<pre> Heuristic addHiddenQtyCond Participants:   ?s CurrentState   ?p ProcessInstance   ?t ProcessType Constraints:   (startsAfterEndingOf ?s ?p)   (isa ?p ?t) Consequences:   (exists ?cq)   (isa ?cq ConceptualQuantity)   (revise ?t (addQtyCondition ?cq)) </pre>	<pre> ModelFragment m2 Participants:   ?e Entity Constraints:   nil Conditions:   (&gt; (q ?e) 0) Consequences:   (&gt; (Rate ?self) 0)   (i- (x-pos ?e) (Rate ?self)) </pre>
--	---

**Figure 49: Left: a heuristic `addHiddenQtyCond` that revises process models by adding a hidden (conceptual) quantity.**

**Right:  $m_2$ , the result of revising  $m_1$  (Figure 48, right) with `addHiddenQtyCond`. Hidden quantity  $q$ , a placeholder force-like quantity, is revisable by other heuristics.**

is applicable when some process model (e.g.,  $m_1$ ) ends before the current state. The consequences of the heuristic (1) assert the existence of a new, hidden quantity  $?_{CQ}$  and (2) revise the conditions of the model fragment ( $m_1$ ) to require the existence of  $?_{CQ}$ . Consider that the system generated the ground symbol  $q$  to represent the conceptual quantity  $?_{CQ}$ . The result is model fragment  $m_2$ , which describes things moving when they have  $q$  at a rate qualitatively proportional to their  $q$ . Hidden *conceptual quantities*, such as  $q$ , are categories that are not observable in a scenario, and their existence is inferred via the conditions and consequences model fragments. The network after applying the heuristic `addHiddenQtyCond` and explaining the ball stopping is shown in Figure 50(b). This includes the new quantity  $q$  and a preference  $m_1 <_c m_2$ . Note that the previous model  $m_1$  still exists in the system – instead of directly revising the model fragment  $m_1$  into  $m_2$ , the system copies  $m_1$  before performing the revision. This copy-revise-prefer approach means that the structure of any previous explanations that use  $m_1$  would remain intact. The preference over model fragments  $m_1 <_c m_2$  causes the derivation of explanation-level preference  $x_0 <_{xp} x_1$ , as described in section 4.6.1. The preference  $m_1 <_c m_2$  also indicates opportunities for retrospective explanation, as discussed in section 4.7. We discuss the role of retrospective explanation later in this chapter.

As noted by Kass (1994), adaptation mechanisms – such as these revision heuristics – fall on a spectrum from (1) a multitude of domain-dependent adaptation strategies, and (2) a smaller number of very general, domain-independent strategies. In this simulation, no heuristic explicitly mentions movement or x/y coordinate quantities, so they are not purely domain-dependent; however, in the case of heuristic `addHiddenQtyCond` (Figure 49) and others like it, heuristics can be very specialized in their applicability. We next discuss how the system chooses between heuristics when several are applicable.

### 8.2.2 Choosing among applicable heuristics

A heuristic's applicability to a situation is determined by its participants and constraints, and its complexity of change is determined by its consequences. Some consequences create new model fragments and categories altogether. For instance, `createDecreaseProcess` created model fragment  $m_I$ , and `addHiddenQtyCond` created a new conceptual quantity  $q$ . These consequences extend the domain knowledge of the agent. Other consequences revise existing model fragments (e.g., `addHiddenQtyCond` revises a model fragment to extend its conditions and consequences) and categories. Heuristics are ordered from minimum to maximum estimated change by tallying their consequences. The cost of each consequence is as follows:

- Revising a conceptual quantity's specification: 3
- Revising (i.e., copying and revising) a model fragment: 7
- Creating an altogether new model fragment: 20
- Creating an altogether new conceptual quantity: 20

Using this cost metric, the system can assign a numerical cost to each applicable heuristic by summing the cost of its consequences. The system then sorts heuristics by ascending cost and executes them in that order until it can explain the situation.

### 8.2.3 Revising conceptual quantities

Like model fragments, conceptual quantities such as  $q$  can be revised using heuristics when the system fails to explain an explanandum. When the quantity  $q$  is created by the heuristic

<pre> Heuristic vectorizeQty Participants:   ?obj Entity   ?quant SpatialQuantity   ?c-quant ConceptualQuantity   ?t ProcessType Constraints:   (increasing (?quant ?obj))   (consequence ?t (i- (?quant ?ent) (Rate ?self)))   (condition ?t (&gt; (?c-quant ?ent) 0)) Consequences:   (isa ?c-quant VectorQuantity)   (revise ?t (addParticipant ?d Direction))   (revise ?t (directionalizeQuantity ?quant ?d))   (revise ?t (directionalizeQuantity ?c-quant ?d)) </pre>	<pre> ModelFragment m<sub>3</sub> Participants:   ?e Entity   ?d Direction Constraints:   nil Conditions:   (&gt; (q[?d] ?e) 0) Consequences:   (&gt; (Rate ?self) 0)   (i+ (pos[?d] ?e) (Rate ?self)) </pre>
--	---

**Figure 51: Left: a heuristic `vectorizeQty` that transforms a scalar conceptual quantity into a vector quantity and revises the according model fragment to take a direction.**

**Right:  $m_3$ , the result of revising  $m_2$  (Figure 49, right) with `vectorizeQty`.**

`addHiddenQtyCond`, it has a magnitude that permits leftward movement. While accommodating subsequent training data, heuristics can revise  $q$  to: (1) add a vector component so that  $q$  has a spatial direction as well as a magnitude; (2) add an influence from another quantity, such that an object's size influences its amount of  $q$ ; (3) add direct influences from process rates, e.g., to describe the transfer of  $q$  between objects or consumption of  $q$ ; or (4) change  $q$  to a quantity that only exists between – and not within – objects. When a quantity  $q$  is revised as  $q'$ , the former specification  $q$  remains, so that existing explanations that use  $q$  are not affected. As with revised model fragments, a preference  $q <_c q'$  is automatically encoded in the network.

To illustrate another failure-based revision, consider an example where the system must explain a cup sliding to the right along a table, but at present, it only has a model fragment describing leftward movement  $m_2$  (Figure 49, right). Rather than construct a new model fragment altogether, it can use the heuristic `vectorizeQty` (Figure 51, left) to revise its model fragment  $m_2$  into model fragment  $m_3$  (Figure 51, right). This heuristic revises both the model fragment as well as the conceptual quantity  $q$ . The quantity  $q$  now has a directional component

such as `left` or `right`, and according to model fragment  $m_3$ , something moves left or right when it has  $q$  in that direction. The symbol `zero` is used to represent the directional component when a quantity is not changing.

### 8.2.3.1 Ontological properties of conceptual quantities

We have described the mechanism by which conceptual quantities are revised, but there are ontological questions regarding the initial conceptual quantity  $q$ . For instance, does  $q$  have a spatial extent? How does it combine with the  $q$  of other objects? How is it acquired or consumed? How does it change its directional component? We look to the cognitive psychology literature for insight.

Pfundt and Duit (1991) analyzed approximately 2,000 published articles about novice misconceptions in the domain of force dynamics. These illustrate that novices do not generally conceive of force as an interaction between two material objects. The most common misconception is that force is a property of a single object. Chi and colleagues (Chi, 2008; Reiner et al., 2000; Chi et al., 1994b) argue that novices often attribute this internal property of force with the ontological properties of a *substance schema*. The substance schema listed in Reiner et al. (2000) contains eleven ontological attributes, though these are not claimed to be complete or globally coherent:

1. Substances are *pushable* (able to push and be pushed).
2. Substances are *frictional* (experience “drag” when moving in contact with a surface).
3. Substances are *containable* (able to be contained by something).
4. Substances are *consumable* (able to be “used up”).

5. Substances are *locational* (have a definite location).
6. Substances are *transitional* (able to move or be moved).
7. Substances are *stable* (do not spontaneously appear or disappear).
8. Substance can be of a *corpuscular nature* (have surface and volume).
9. Substances are *additive* (can be combined to increase mass and volume).
10. Substances are *inertial* (require a force to accelerate).
11. Substances are *gravity sensitive* (fall downward when dropped).

Not all of these attributes are relevant for our system's conceptual quantity  $q$  that mediates motion, but we use these guidelines for constraining the properties of conceptual quantities when there is any question. For instance, according to model fragment  $m_2$ ,  $q$  is a property of an entity and not an abstract property, so it is *locational* and (in some sense) *containable*. Additionally, since conceptual quantities must be *stable*, the system must justify how an object's  $q$  increases and decreases. This is achieved using: (1) processes that describe consumption of quantities over time so that  $q$  is *consumable* (see also the "dying away" p-prim in diSessa, 1993) and (2) processes that describe the transfer of quantities between objects so that  $q$  is *transitional*.

If we apply these principles to the directional conceptual quantity  $q$  described within model fragment  $m_3$ , an entity has  $q[\text{left}]$  when it travels leftward,  $q[\text{right}]$  when it travels rightward, and  $q[\text{zero}]$  when it is still. This means that other processes affect the direction of an object's  $q$ , which in turn affects the object's position in space. Without a transfer across objects, the sum of an object's  $q$  across directions is constant. This satisfies the stability constraint of the substance schema.

In the psychology literature, Ioannides and Vosniadou (2002) are investigating the “meaning” of force. In our model, the meaning of a quantity (e.g.,  $q$ ) is a conjunction of these ontological constraints on the quantity, the direct and indirect influences, and model fragments (e.g.,  $m_2$  in Figure 49) that describe the existence and behavior of the quantity within a scenario. As quantities and model fragments change, so will the presence and role of  $q$  within the questionnaire scenarios that we use as testing data.

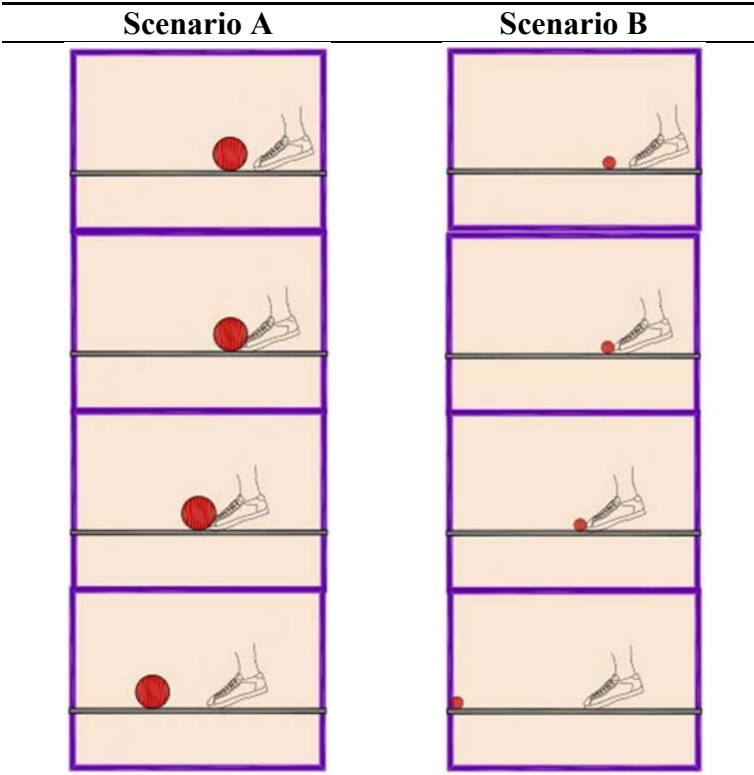
Thus far, we have described how the system explains quantity changes within observations and revises model fragments and quantities. However, the system also explains differences in behavior *between* similar observations, using analogy. This comparative explanation process is important for finding qualitative proportionalities between quantities. We discuss this next.

#### 8.2.4 Inter-scenario analysis

After the system explains the quantity changes within a comic graph observation, it retrieves a similar previous observation to determine whether there are any discrepancies. If there are variations in the quantity changes between observations (e.g., one object moves further than another object) then they must be explained. Failure to explain these discrepancies results in the use of heuristics to revise domain knowledge, as described above. We call this *inter-scenario analysis*, and we illustrate this with an example.

Suppose that the comic graph labeled “Scenario A” in Figure 52 has already been explained by the simulation. Suppose also that the simulation has just explained the quantity changes within a second comic graph labeled “Scenario B” in Figure 52. Scenario B is identical to Scenario A, except that a *smaller* ball is kicked a *greater* distance. Finally, suppose that both scenarios were explained using the same specification of  $q$  and model fragment  $m_3$  (Figure 51,

right), as well as process model fragment instances that describe the  $q[\text{zero}]$  of an object transitioning to  $q[\text{left/right}]$  of the object, and visa-versa, obeying the stability constraint of the substance schema.



**Figure 52: Comic graph scenarios A and B are sufficiently similar for inter-scenario analysis.**

After the system explains Scenario B, it uses MAC/FAC to retrieve a similar, previously-explained scenario. If the SME normalized similarity score<sup>42</sup> between the probe and the previous scenario is above a threshold value (we use 0.95 in our simulation), then inter-scenario analysis proceeds between the two scenarios. Suppose that Scenario A is retrieved, and that the SME mapping between Scenarios A and B exceeds the similarity threshold.

Inter-scenario analysis between Scenarios A and B involves explaining why corresponding quantities changed differently in Scenario A than they did in Scenario B, if applicable. For

<sup>42</sup> See section 3.4.1 for a description of how this is computed.



instance, the ball in Scenario B travels a greater distance along the x-axis than the ball in Scenario A. These quantity change variations are detected by analyzing correspondences in the SME mapping between Scenarios A and B, several of which are shown in Figure 53. From these correspondences, the system can compute two inequalities, shown in the right column of Figure 53:

$$\begin{aligned} (\text{Area ball-a}) &> (\text{Area ball-b}) \\ (\Delta x[\text{left}] \text{ ball-a}) &< (\Delta x[\text{left}] \text{ ball-b}) \end{aligned}$$

The inequality between quantity changes  $(\Delta x[\text{left}] \text{ ball-a}) < (\Delta x[\text{left}] \text{ ball-b})$  must be explained. As above, heuristics are used to revise knowledge to aid in explanation.

Scenario A formula	Scenario B formula	Inequality (if applicable)
foot-a	foot-b	<i>n/a</i>
ground-a	ground-b	<i>n/a</i>
ball-a	ball-b	<i>n/a</i>
mfi-a	mfi-b	<i>n/a</i>
(isa mfi-a $m_3$ )	(isa mfi-b $m_3$ )	<i>n/a</i>
(Area ball-a)	(Area ball-b)	$(\text{Area ball-a}) > (\text{Area ball-b})$
$(\Delta x[\text{left}] \text{ ball-a})$	$(\Delta x[\text{left}] \text{ ball-b})$	$(\Delta x[\text{left}] \text{ ball-a}) < (\Delta x[\text{left}] \text{ ball-b})$
$(q[\text{left}] \text{ ball-a})$	$(q[\text{left}] \text{ ball-b})$	$(q[\text{left}] \text{ ball-a}) ? (q[\text{left}] \text{ ball-b})$
(Rate mfi-a)	(Rate mfi-b)	$(\text{Rate mfi-a}) ? (\text{Rate mfi-b})$
$(i+ (x\text{-pos}[\text{left}] \text{ ball-a})$ $(\text{Rate mfi-a}))$	$(i+ (x\text{-pos}[\text{left}] \text{ ball-b})$ $(\text{Rate mfi-b}))$	<i>n/a</i>
$(> (q[\text{left}] \text{ ball-a}) 0)$	$(> (q[\text{left}] \text{ ball-b}) 0)$	<i>n/a</i>
...	...	...

**Figure 53: Selected analogical correspondences between Scenarios A and B (Figure 52).**

The first task in explaining the quantity change inequality is to derive other inequalities between corresponding quantities. As mentioned above, the movements of ball-a and ball-b were explained using model  $m_3$  in Figure 51(right). Since the  $m_3$  model fragment instances mfi-

a and mfi-b of Scenarios A and B correspond (see Figure 53), so do their respective rates, (Rate mfi-a) and (Rate mfi-b), and their respective conditions ( $(q[\text{left}] \text{ ball-a}) > 0$ ) and ( $(q[\text{left}] \text{ ball-b}) > 0$ ). The corresponding conditions are especially important, since (1) conditions must hold for the processes mfi-a and mfi-b to be active and (2) assuming a closed world, these processes are the only influences of  $\Delta x[\text{left}]$  for both balls. If we assume the rates of the processes are the same,<sup>43</sup> the variation in  $\Delta x[\text{left}]$  between ball-a and ball-b is a factor of the  $(q[\text{left}] \text{ ball-a})$  and  $(q[\text{left}] \text{ ball-b})$ , which varied the duration of these process instances. This produces the following ordinal relation to describe the relative  $q$  values in the transition to last frame of the comic graphs:

$$(q[\text{left}] \text{ ball-a}) < (q[\text{left}] \text{ ball-b})$$

The variation in  $\Delta x[\text{left}]$  has been explained with an inequality in  $q[\text{left}]$  values in this state, but now the inequality between  $q[\text{left}]$  values in this state requires an explanation. This will require that the system revises its beliefs about  $q$ . Since  $q$  is a conceptual quantity created by the system, there are many ways to explain this inequality. We use the substance schema of Reiner et al. (2000) to constrain the system's explanation. The following inferences are plausible with respect to the substance schema:

1. ball-b has more total  $q$  than ball-a in the movement states, but this is consumed before the resting state is reached.

---

<sup>43</sup> The system's explanation of this quantity variation relies on the assumptions the system makes about time. For instance, if we assume the transitions between corresponding frames in Scenarios A and B take equal time, then the variation in leftward movement can only be explained by varying rates of change. If we do not make this assumption, we can explain variation of leftward movement with equal rates of change and one process being active longer than the other, corresponding process.

2. ball-b has more  $q[\text{left}]$  than ball-a in the movement states, but this transitions to  $q[\text{zero}]$  before the resting state.

Inference (1) is not plausible in our example, since the simulation does not have a model of how a greater total amount of the conceptual quantity  $q$  is initially acquired by the ball. In simulation trials where the system constructs a model of  $q$  transfer prior to this analysis, this is the path chosen by the system. Inference (2) obeys the stability constraint of the substance schema as well as the present properties of the conceptual quantity. As a result, the system must explain why ball-b has greater  $q[\text{left}]$  and less  $q[\text{zero}]$  in the movement state. This is done by asserting a new qualitative proportionality to another varying quantity. In this case, the inequality  $(\text{Area ball-a}) > (\text{Area ball-b})$  is used, since it is the only other varying quantity. The system asserts the following statements for entities  $?ent$  and directions  $?dir$ :

```

if ?dir  $\neq$  zero:

    (qprop- (q[?dir] ?ent) (Area ?ent))

    (qprop (q[zero] ?ent) (Area ?ent))

```

This states that all else being equal, smaller objects have more directional (e.g., left or right)  $q$ , which propels them further than larger objects. Larger objects have greater  $q$  in the zero direction. If a second quantity, such as the size of the foot, varied in addition to the size of the balls, neither would be isolated. Consequently, either or both might explain the variation in  $\Delta x[\text{left}]$ , and inter-scenario analysis terminates without revising the quantity. This makes the

simulation more conservative when hypothesizing qualitative proportionalities, since it requires pairwise quantity variations in isolation.

After this conceptual quantity is revised by adding the above qualitative proportionality, inter-scenario analysis is complete. When the simulation uses the revised quantity  $q$  and the associated model fragments to answer the questionnaire, it will assert that entities have exactly one  $q$  property, and its magnitude is a function of its size. These answers are consistent with the “Internal” meaning of force according to the coding scheme of Ioannides and Vosniadou (2002).

### **8.2.5 Retrospective explanation propagates revisions**

We have described how the simulation revises its knowledge when it fails to explain observations or when it fails to explain variations between similar observations. Instead of revising a construct (i.e., model fragment or quantity) directly, the system copies it and then revises the copy so that the prior construct remains. The agent then encodes an epistemic preference for the new construct over the prior one. Figure 50 illustrates this copy-revise-prefer behavior. After the revision, quantity changes that were explained with the prior construct retain their present explanations, despite the fact that these explanations rely on outdated domain knowledge.

The process of retrospective explanation, described in section 4.7, constructs new explanations to replace these outdated explanations. Retrospective explanation is achieved through the following steps in this simulation:

1. Find an outdated explanandum (i.e., quantity change)  $m$ . An explanandum is outdated if and only if (1) there is a concept-level preference  $b <_c b'$  and (2) the preferred explanation for  $m$  is  $x$ , and  $x$  uses  $b$  and not  $b'$ .
2. Attempt to explain the explanandum with preferred knowledge, using the same explanation construction algorithms as above.
3. Compute preferences over new explanations and previous explanations, using the same explanation evaluation algorithms as above.
4. Map the explanandum to a new, preferred explanation, if applicable.
5. If the outdated explanandum  $m$  still retains its previously preferred explanation, store the triple  $\langle m, b, b' \rangle$  so that this process is not later repeated for the same purpose.

Retrospective explanation is an incremental transition from one causal description to another.

This models the students' incremental transition to a new understanding of the world.<sup>44</sup>

In this simulation, retrospective explanation occurs to completion after each new training datum is given to the system. This means that every local revision to domain knowledge is immediately used to explain previous observations.

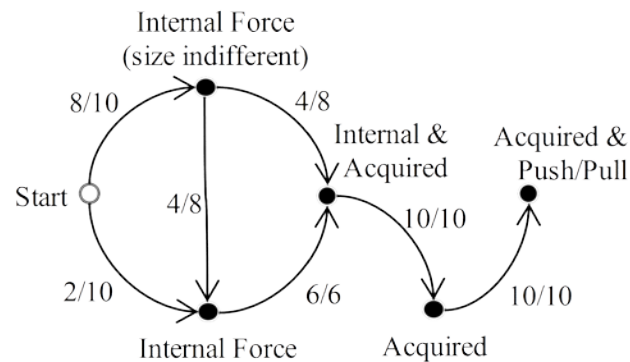
### 8.3 Simulation results

Here we describe the setup and results of our simulation. The psychological assumptions and justification of our match with student data is addressed in section after this.

---

<sup>44</sup> Following McDermott (1976), this is not to suggest that the simulation is itself “understanding” the phenomena.

We used ten comic graphs as training data, for a total of 58 comic graph frames and 22 instances of movement. These were sketched in CogSketch. The system was also given starting knowledge about agency, such that people and their respective body-parts cause their own translation. The system uses this knowledge to explain how a person or body-part (1) starts or stops translating or (2) imparts or consumes a conceptual quantity (e.g.,  $q$ , in the above examples) if that quantity causes movement. Modeling how agency and intentional movement is learned is a nontrivial and interesting research problem, but is beyond the scope of this simulation.



**Figure 54: Changes in the simulation's meaning of force, using Ioannides and Vosniadou's (2002) student meanings of force.**

For each comic graph used as a training datum, the system: (1) explains all quantity changes within the comic graph; (2) retrieves a similar previous comic graph using MAC/FAC, using the present one as a probe; (3) performs inter-scenario analysis if the present and previous comic graphs have a SME normalized similarity score above 0.95; and (3) performs complete retroactive explanation if model fragments or quantities were revised.

After a comic graph is processed in this manner, the system completes the entire questionnaire, half of which is shown in Figure 47. From the system's answers, we determine (1) the conditions under which a force-like quantity exists, and (2) the effect of factors such as

size, height, and other agents on the force-like quantity. We use the same coding strategy as Ioannides & Vosniadou (2002) to determine which meaning of force the system has learned, given its answers on the questionnaire. No knowledge revision occurs during the system's completion of the questionnaire.

Figure 54 illustrates the transitions in the concept of force across 10 independent trials with different comic graph order. The simulation starts without any process models or quantities to represent force, and transitions to the “Internal Force” concept 2/10 times, and a size indifferent “Internal Force” model 8/10 times, which was not reported by Ioannides & Vosniadou (2002). In these cases, the force-like quantity (e.g.,  $q$ ) was not qualitatively proportional to size. The rest of the transitions follow a similar trajectory to the student data in Figure 46. Each trial of the simulation completes an average of six model fragment revisions and four category revisions of a placeholder force-like quantity during its learning.

## 8.4 Discussion

We have simulated the incremental revision of a force-like category over a sequence of observations. As it incorporates new observations, the system occasionally fails to explain (1) quantity changes within the observation and (2) why quantity changes vary between similar observations. In response to these anomalous situations, the system minimally and incrementally revises its domain knowledge using declarative heuristics. It then propagates these local revisions to other contexts via retrospective explanation.

Human conceptual change in the domain of force dynamics occurs over a span of years for the students in Ioannides and Vosniadou (2002). This can be inferred from Figure 46, though the data at each age group were gathered from different students. Over these years, students are

exposed to a multitude of observations, formal instruction, and physical interaction. Providing this amount of input and a similarly varied workload is beyond the state of the art in cognitive simulation. Consequently, this simulation learns via sketched observations alone, so the set of stimuli is smaller and much more refined. Since the knowledge encoded from CogSketch is not as rich as human perception, our simulation relies upon the psychological assumptions stated in the CogSketch discussion in Chapter 3. Before revisiting the hypotheses of this simulation, we discuss factors that enabled the simulation to transform its domain knowledge so rapidly. These factors involve the training data and the computational model itself.

Since comic graphs are already segmented into qualitative states, the system does not have to find the often-fuzzy boundaries between physical behaviors. Furthermore, the sketches convey relative changes in position, but not relative changes in velocity, so the system needs not differentiate velocity from acceleration, which is difficult for novice students (Dykstra et al., 1992). Finally, the comic graphs are sparse, which simplifies the detection of anomalies. Some are also highly analogous (see Figure 52), which facilitates inter-scenario analysis.

Aside from the comic graph stimuli, aspects of the computational model itself accelerate learning beyond human performance. People have many strategies they can use to discredit anomalous data (Feltovich et al., 2001), and other tactics to avoid conceptual change, such as explaining away, excluding anomalous data from theories, reinterpreting anomalous data to fit within a theory, holding data in abeyance, and making partial or incomplete changes (e.g., Chinn & Brewer, 1998). In fact, complete conceptual change is actually a last-resort for children (Chinn & Brewer, 1998). Our system's sole response to any explanation failure is the revision of domain knowledge, followed by exhaustive retrospective explanation. Our intent in this simulation is to model a trajectory of minimal – yet successful – conceptual changes. While



modeling these conceptual change avoidance strategies is beyond the scope of the present simulation, it is an interesting opportunity for future work, and we revisit this idea in Chapter 9.

A final possible cause for the simulation's accelerated learning is the heuristics used by the simulation. The mechanisms by which students spontaneously revise their domain knowledge are unknown. As discussed in Chapter 2, there is considerable debate regarding how such knowledge is even represented and organized. The heuristics used in this simulation may skip intermediate steps, and thereby make larger changes than people spontaneously make to their mental models and categories. Alternatively, the heuristics used in this simulation may revise domain knowledge in altogether different fashions than children do upon explanation failure. For example, conceiving of force as an interaction between objects (e.g., “Push/Pull” and “Acquired & Push/Pull” meanings) may be the result of social interaction and reading (e.g., the familiar sentence “A force is push or a pull”) and not of error-based revision.

The three trajectories (i.e., unique paths through the graph) illustrated in Figure 54 describe plausible paths through the human data in Figure 46, supporting the hypothesis that our explanation-based framework can simulate human category revision. The most popular – but still incorrect – category of force “Gravity & Other” is not reached by the simulation. This category requires the mention of *gravity*, which is not learned by the simulation, and is almost certainly learned by students through formal instruction and social interaction.

This simulation supports the hypothesis that compositional model fragments can simulate the mental models of the students in this domain. Compositional models are used here to infer the presence of unobservable, force-like quantities within scenarios. The system infers the presence and relative magnitudes of these quantities in a fashion comparable with students, and

is able to simulate multiple student misconceptions on the same questionnaire. This supports the knowledge representation hypothesis.

## Chapter 9: Conclusion

“Nothing endures but change”

– Heraclitus

We have described a computational model of conceptual change and used it to simulate results from the literature on conceptual change in students in different domains. Chapter 1 presented the claims of the dissertation, an outline of our model, and its psychological assumptions. Chapter 2 discussed four existing theories of conceptual change and areas of disagreement between them to identify where our model could shed some light. Chapter 3 reviewed the AI techniques used in the computational model, and Chapter 4 presented the computational model itself. The computational model was used to perform four simulations, described in Chapters 5-8, providing empirical evidence to support the claims of this dissertation.

This chapter revisits our claims in light of the evidence provided by the simulations. We then discuss related AI systems and compare our computational model to the other theories of conceptual change described in Chapter 2. We close with a discussion of general limitations and opportunities for future work.

### 9.1 Revisiting the claims

Here we discuss each claim of this dissertation. The first claim is about knowledge representation:

***Claim 1:*** Compositional qualitative models provide a psychologically plausible computational account of human mental models.

Claim 1 is not a new idea, since simulating human mental models was an early motivation for qualitative modeling in AI (Forbus & Gentner, 1997). However, the simulations described in Chapters 5-8 offer novel evidence in support of this claim. Since this is a knowledge representation claim, it we supported it by (1) observing how people construct explanations and solve problems with their mental models from the cognitive science literature and (2) using compositional qualitative models to construct the same explanations and solve the same problems. We used qualitative models in all four simulations, to simulate student problem-solving in three domains:

1. Force dynamics (Chapters 5 and 8)
2. Astronomy (Chapter 6)
3. Biology (Chapter 7)

In addition, our system used the qualitative models that it learned to perform different problem-solving tasks, with results similar to students:

1. Explaining causal models of a dynamic system (Chapters 6 and 7)
2. Predicting the next state of a scenario (Chapter 5)
3. Explaining abstract events in sketched scenarios (Chapter 5)
4. Explaining hidden mechanisms in sketched scenarios (Chapter 8)

The second claim involves learning by induction:

**Claim 2:** Analogical generalization, as modeled by SAGE, is capable of inducing qualitative models that satisfy Claim 1.

Claim 2 is a novel claim, since AI systems have not previously induced qualitative models using SAGE. Chapter 5 supports this claim with empirical evidence by using sketched observations as training data, inducing qualitative models from these training data, and then using the resulting qualitative models to perform two problem-solving tasks in a fashion consistent with human students.

As we describe in Chapter 5, SAGE does not produce qualitative models directly; rather, SAGE produces probabilistic generalizations of the input observations. The simulation transforms these into qualitative models by (1) filtering out low-probability statements and (2) creating a qualitative model using the temporal data within the remaining high-probability statements. The resulting model describes the participants, preconditions, causes, and effects of events.

The third claim involves modeling two types of conceptual change:

**Claim 3:** Human mental model transformation and category revision can both be modeled by iteratively (1) constructing explanations and (2) using meta-level reasoning to select among competing explanations and revise domain knowledge.

Chapter 4 described how explanations are constructed and how meta-level reasoning decides which explanation is preferred, when multiple explanations apply. When the model replaces its

preferred explanation for a phenomenon (e.g., how blood flows from the heart to the body) with a new explanation, it will use the beliefs within the new explanation to solve similar problems and answer related questions in the future. This means that replacing a preferred explanation is a context-sensitive revision of beliefs. The simulations in Chapters 6, 7, and 8 exemplify this behavior. In these three simulations, a sequence of these revisions simulate the adoption of new causal mechanisms (Chapter 6), the integration of new components into an existing mental model (Chapter 7), and the transformation of a category (Chapter 8).

The third claim also mentions meta-level revision. In Chapter 8, the system copies and revises its domain knowledge when it fails to consistently explain a phenomenon. By using declarative heuristics, the model can estimate the amount of change a heuristic will incur to domain knowledge and then choose the one that incurs the least estimated change. This revision operation frees the system from a failure mode, so the system then resumes the above explanation construction and explanation evaluation methods.

The simulation results presented here provide evidence that my model is a plausible account of human conceptual change.

## **9.2 Related work in AI**

Here we discuss other AI systems that learn about new quantities, causal mechanisms, and causal relationships between phenomena. Only two of the systems we review, INTHELEX and ToRQUE2, have been used to simulate human conceptual change. Since the rest of these systems are not cognitive models, we compare them to our model in terms of the knowledge representations and algorithms used, since there are relevant overlaps.

The Qualitative Learner of Action and Perception (QLAP) (Mugan & Kuipers, 2011) learns hierarchical actions from continuous quantities in an environment. QLAP uses qualitative reasoning to discretize continuous quantities into intervals, using the quantities' landmark values. Dynamic Bayesian networks (DBNs) are then used over open intervals and values in each quantity space to track contingencies between qualitative values and events in the world. This is useful for learning preconditions for events in a continuous world. This could provide an account for how preconceptions might be learned from experience, but does not account for how they are revised by instruction or explanation failures.

Automated Mathematician (AM) (Lenat & Brown, 1984) was an automated discovery system that used heuristics to apply and revise domain knowledge. AM operated within the domain of mathematics, with its concepts represented as small Lisp programs. The control structure involved selecting a mathematical task from the agenda and carrying it out with the help of heuristics that activate, extend, and revise AM's mathematical concepts. The mathematical concepts were then used for solving problems on AM's agenda. EURISKO (Lenat, 1983) improved upon AM by using a more constrained frame-based representation and allowing heuristics to modify other heuristics. This provided a more sophisticated meta-level, where components influenced each other in addition to the mathematical concepts. Both AM and EURISKO contained structures designed to control and mutate the object-level concepts that did the primary domain-level reasoning. Also, both systems relied on humans to evaluate the intermediate products of reasoning, where our model learns autonomously from instruction and observation. Additionally, our model incorporates other types of reasoning such as analogy, abduction, and qualitative reasoning to learn in scientific domains.

Meta-AQUA (Ram & Cox, 1994; Cox & Ram, 1999) is a story understanding system that learns from expectation failures. The system monitors its progress in explaining events within stories. When explanation fails, it triggers meta-level control to set knowledge goals such as reorganizing hierarchies and acquiring new information. It does this using two general representations for metareasoning: (1) *Meta-XPs*, which describe the system's goal-directed reasoning, and (2) *Introspective Meta-XPs*, which describe a failure in reasoning, rationale for the failure, the knowledge goals to solve the failure, and algorithms for satisfying the knowledge goals. Like our category revision simulation in Chapter 8, Meta-AQUA uses metareasoning in reaction to failure by identifying deficits in knowledge and proposing repairs.

ECHO (Thagard, 2000) is a connectionist model that uses constraint satisfaction to judge hypotheses by their explanatory coherence. This is designed to model how people might revise their beliefs, given the propositions and justification structure in their working memory. ECHO operates at the level of propositions, creating excitatory and inhibitory links between consistent and inconsistent propositions, respectively. ECHO uses a winner-take-all network, which, while computationally powerful, means that it cannot distinguish between absence evidence for competing propositions versus balanced conflicting evidence for them. ECHO does not generate its own theories or justification structure, as our system does.

ACCEPTER (Leake, 1992; Schank et al., 1994) is a case-based reasoning system that detects anomalies within a situation and resolves them by constructing explanations. After detecting an anomaly, ACCEPTER encodes an *anomaly characterization* that sets knowledge goals and helps retrieve relevant explanation patterns (Schank, 1986) from a library thereof. It then evaluates candidate explanation patterns with respect to whether it explains the anomaly, and whether it is plausible. For instance, explaining the *Challenger* explosion as a Russian



sabotage is implausible because Russia would not risk a dangerous confrontation with the United States. As in our model, constructing and evaluating explanations is central to ACCEPTER; however, our system also replaces explanations with preferred ones to perform belief revision.

One problem of case-based explanation systems such as ACCEPTER is that retrieved cases and explanation patterns may not apply to the present context. When TWEAKER (Schank et al., 1994) retrieves an explanation that is a close – but not perfect – match to the current problem, it uses *adaptation strategies* to build variations of the explanation. These adaptations include replacing an agent, generalizing or specifying slot-fillers, and so-forth. TWEAKER can also use *strategy selection* to choose between possible strategies, which helps guide search through a large explanation search space. Our category revision simulation in Chapter 8 is similar to TWEAKER in that it uses revision heuristics as its adaptation strategies, and it scores and sorts heuristics as its strategy selection.

INTHELEX (Esposito et al., 2000) is an incremental theory revision program that has modeled conceptual change as supervised learning. It implements belief revision as theory refinement, so it minimally revises its logical theories whenever it encounters an inconsistency. INTHELEX is capable of learning several intuitive theories of force from observations, but it has not simulated the transition from one intuitive theory to another. The transition between intuitive theories (e.g., in Chapter 8) is a central principle for simulating conceptual change, so while INTHELEX may simulate how intuitive theories are acquired, it does not simulate conceptual change at the scale proposed in this dissertation.

The ToRQUE and ToRQUE2 systems (Griffith, Nersessian, & Goel, 1996; 2000) solve problems using *structure-behavior-function* (SBF) models. To solve a new target problem, ToRQUE2 retrieves analogs to the present problem, and then applies transformations to the

analog or target problems to reduce their differences. This generates additional SBF models and generates a solution to the target problem using transformed domain knowledge. ToRQUE2 has simulated how scientists solve problems during think-aloud protocols, where the scientists change their understanding throughout the problem-solving session. For instance, a scientist initially believes that the stretch of a spring is due to its flexibility, and then realizes that a spring maintains constant slope when stretched through torsion in the spring's wire. The authors conclude that this spring example is "an instance of highly creative problem solving leading to conceptual change" (p. 1). Since ToRQUE2 revises domain knowledge to overcome failures in problem-solving, and the new spring model conflicts with the previous one, this is a type of mental model transformation. By comparison, conceptual change is triggered differently in our cognitive model, and our model searches for consistent, low-complexity models that fit multiple observations (e.g., Chicago's and Australia's seasons, in Chapter 6).

Explanation-Based Learning (EBL) systems (DeJong, 1993) learn by creating explanations from existing knowledge. Many EBL systems learn by *chunking* explanation structure into a single rule (e.g., Laird et al., 1987). Chunking speeds up future reasoning by avoiding extra instantiations when a macro-level rule exists, but it does not change the deductive closure of the knowledge base, and therefore cannot model the repair of incorrect knowledge. Other systems use explanations to repair knowledge. For example (Winston and Rao, 1990) uses explanations to repair error-prone classification criteria, where explanations are trees of if-then rules over concept features. Upon misclassification, the system analyzes its explanations and creates censor rules to prevent future misclassification. Similarly, our model detects inconsistencies within and across explanations in its analysis, but it encodes epistemic preferences (rather than censor rules) to resolve these issues.

Other systems construct explanations using abduction to extend or revise their domain knowledge. Molineaux et al. (2011) describes a system that determines the causes of plan failures through abduction. Abduction increases the agent's knowledge of hidden variables and consequently improves the performance of planning in partially-observable environments. Similarly, ACCEL (Ng & Mooney, 1992) creates multiple explanations via abduction, and it uses simplicity and set-coverage metrics to determine which is best. When performing diagnosis of dynamic systems, ACCEL makes assumptions about the state of components (e.g., a component is abnormal or in a known fault mode), and minimizes the number of assumptions used. By contrast, when our system evaluates explanations, some assumptions (e.g., quantity changes) are more expensive than others, and other artifacts (e.g., contradictions, model fragments, and model fragment instances) incur costs.

Other systems reason with abduction under uncertainty while still using structured relational knowledge. Bayesian Abductive Logic Programs (Raghavan & Mooney, 2010) and Markov Logic Networks (Richardson & Domingos, 2006; Singla & Mooney, 2011) have been used for these purposes. Uncertainty is an important consideration for reasoning about psychological causality (e.g., recognizing an agent's intent) and for reasoning about physical phenomena in the absence of mechanism-based knowledge. In this thesis we are specifically concerned with abduction using mechanism-based knowledge, so probability distributions are not as central as for other tasks and domains. That said, probabilities might represent the agent's purported likelihood of a given belief or model fragment in one of the domains simulated here, which could direct the search for explanations. We revisit this idea below.

Previous research in AI has produced postulates for belief revision in response to observations. The AGM postulates (Alchourrón et al., 1985) describe properties of rational

revision operations for expansion, revision, and contraction of propositional beliefs within a deductively-closed knowledge base. Katsuno and Mendelzon's (1991) theorem shows that these postulates can be satisfied by a revision mechanism based on total pre-orders over prospective KB interpretations. Like these approaches, our conceptual change model computes total pre-orders over belief sets, but our system is concerned with consistency within and across preferred explanations rather than within the entire KB. Further, since our model has an explanatory basis, it uses truth maintenance methods (Forbus & de Kleer, 1993) to track the justification structure and assumptions supporting its beliefs.

### **9.3 Comparison to other theories of conceptual change**

Our computational model shares some psychological assumptions with individual theories of conceptual change discussed in Chapter 2. We review important overlaps and disagreements with each theory, citing examples from our simulations to illustrate.

#### **9.3.1 Knowledge in pieces**

Like the knowledge in pieces perspective (diSessa, 1988; 1993; diSessa et al., 2004), our computational model assumes that domain knowledge is – at some level – stored as individual elements. These elements are combined into larger aggregates to predict and explain phenomena, and can then be recombined into new constructs to accommodate new information. Additionally, when new information is encountered via observation or formal instruction, the new information coexists with the previous elements, even when they are mutually incoherent or inconsistent.

Our theory diverges from knowledge in pieces regarding the structures that organize these domain knowledge elements and the representation of the elements themselves. In our model, explanations are persistent structures that aggregate domain knowledge elements. The knowledge in preferred explanations is reused to explain new observations and solve new problems. Belief revision is performed by revising which explanations are preferred, which thereby affects future reuse of knowledge. By contrast, knowledge in pieces assumes a set of structured cueing priorities that activate these elements in working memory, based on how these elements were previously coordinated (diSessa, 1993). Belief revision is achieved by altering these priorities. Additionally, knowledge in pieces assumes several types of domain knowledge, including p-prims, propositional beliefs, causal nets, and coordination classes. By contrast, our model uses only propositional beliefs and model fragments.

### **9.3.2 Carey's theory**

Like Carey's (2009) theory of conceptual change, our computational model assumes that a single category such as *force* can have multiple, incommensurable meanings. The student has simultaneous access to both of these meanings, but they are contextualized. In both Carey's theory and our model, conceptual change is driven by these category-level conflicts, but in our model, conceptual change is also driven other explanatory inconsistencies and preferences. Also like Carey's theory, our computational model relies on the processes of analogy, abduction, and model-based reasoning to achieve conceptual change.

Our model differs from Carey's theory on how knowledge is contextualized. Carey (2009) assumes that new conceptual systems are established to store incommensurable categories, and that analogy, abduction, and model-based thought experiments add causal structure to these new

conceptual systems. Our model's knowledge is contextualized at the explanation level, so that two phenomena may be explained using mutually incoherent or inconsistent explanations. When our model finds contradictions across preferred explanations, these are resolved locally, to increase the coherence between these explanations. Thus, our model adopts new information and revises its explanations to improve coherence (e.g., by reducing cost, in Chapter 6), but it does not strongly enforce coherence in a discrete conceptual system.

### **9.3.3 Chi's categorical shift**

Like Chi's (2008; 2000) theory of conceptual change and mental model transformation, our computational model relies on self-directed explanation to integrate new information. Chi's (2008) account of mental model transformation involves a series of belief-level refutations, which cause belief revision and the adoption of instructional material. These belief revisions change the structure, assumptions, and predictions of a mental model. In our system, the model of a system such as the human circulatory system is comprised of model fragments and propositional beliefs. As in Chi's theory, revising propositional beliefs can change the structure of this model.

Our model differs from Chi's theory in how it revises information. Chi (2008) assumes that categories are directly shifted across ontological categories, e.g., the category "force" is shifted from a "substance" to a "constraint-based interaction." The category is only shifted once the target category (i.e., "constraint-based interaction") is understood. The number of resulting changes to ontological properties and the unfamiliarity of the target category both increase the difficulty of the change. By contrast, the simulation in Chapter 8 uses heuristics to revise the properties of a category, and the new and old categories coexist, albeit in different explanations.

The system incrementally transitions to the new category by a process of retrospective explanation.

### **9.3.4 Vosniadou's framework theory**

Vosniadou's (1994; 2002; 2007) theory of conceptual change assumes that students have a generally coherent framework theory. The framework theory consists of specific theories about phenomena, mental models of systems and objects, and presupposition beliefs that constrain the theories and mental models within the framework. Our model has similar interdependencies between constructs, but these are soft constraints. For example, in Chapter 6, the system was given the credible information that Australia and Chicago experience opposite seasons at the same time. This information in adopted domain knowledge constrained the explanations of Australia's and Chicago's seasons. In this manner, credible beliefs in adopted domain knowledge are analogous to presuppositions, and specific theories are analogous to explanations.

One important difference between Vosniadou's theory and our model is that Vosniadou's theory assumes a generally coherent framework theory, where our model utilizes local explanatory structures. In our model, coherence and consistency are secondary, macro-level phenomena; they are not hard requirements on the system of beliefs. Our model holds internally-consistent, globally-inconsistent explanations in memory simultaneously and then increases global coherence using cost-based belief revision (e.g., in the seasons simulation in Chapter 6) and retrospective explanation (e.g., in the simulations in Chapters 7 and 8).

Like Vosniadou's theory, our model makes the minimal change to categories such as force (e.g., in Chapter 8) to resolve contradictions. Importantly, the prior category of force ceases to exist in Vosniadou's framework theory because it is inconsistent with the new version of the

category. Conversely, our model retains the prior category and encodes a preference for the revised version. It then incrementally transitions to it via retrospective explanation, when possible (e.g., in Chapter 8). This means that in our model, the agent's knowledge is not globally coherent or even globally consistent. The processes of cost-based belief revision (Chapter 6) and preference-based retrospective explanation (Chapters 7 and 8) make local, incremental repairs to improve adopt preferred knowledge and reduce complexity.

### **9.3.5 Novel aspects of our model as a theory of conceptual change**

As a theory of human conceptual change, our model relies more heavily on the processes of explanation (e.g., Chapters 6-8) and comparison (e.g., Chapters 5 and 8) than these other theories of conceptual change. As discussed in Chapter 1, our model assumes that explanations are persistent structures that organize domain knowledge. Further, it assumes that phenomena are associated with their preferred explanation in memory, so that people can retrieve a previously-explained observation and use its explanation – or the knowledge therein – to explain a new observation using first principles reasoning. The assumption that people retain the complete structure of explanations is probably too strong, and we discuss opportunities for relaxing this assumption below.

In the theories of Chi, Carey and Vosniadou, we can point to a “completed” state of conceptual change. Consider the following examples of completing conceptual change:

- In Chi's theory, consider a student who conceives of “force” as a type of “substance.” She learns a target category such as “constraint-based interaction” (Chi et al., 1994a;



Chi, 2008), and then shifts the concept of “force” to become a subordinate of this target category.

- In Carey’s theory, consider a student who has mistaken knowledge of “force” and “mass” concepts. During formal instruction, a new conceptual system is established to store new categories of “force” and “mass” that are incommensurable with existing categories of the same name. Instruction provides the relational structure between these new symbols, and modeling processes provides and causal structure for the new conceptual system.
- In Vosniadou’s theory, consider that a student believes the earth is flat, like a pancake. She revises set of presuppositions are about the earth, and now conceives of it as an astronomical object. This means that objects on the “sides” and “bottom” of the earth do not fall off. This alters the constraints on her mental models of the earth, so she revises her mental model of the earth to be a sphere, with people living on the “sides” and “bottom” as well.

Is there a similar absolutely “completed” narrative for our model? It seems unlikely. To illustrate, suppose that our model has learned and used a category of force similar to the “Internal” meaning of force (see Chapter 8) to explain many, diverse, phenomena. If it copies and revises this category of force, it can quickly use the revised version to retrospectively explain a very small but salient subset of her experiences. If these experiences are the ones most frequently retrieved for future learning and question answering, the new category and model fragments will be propagated. However, a *completed* conceptual change would require that every observation explained with the prior category is retrospectively explained with the new

category. This seems unlikely. However, it does capture an important property of human mental model reasoning, that people do indeed have multiple, inconsistent models for the same phenomena in different circumstances (Collins & Gentner, 1987).

The absence of a “completed” state in our model means that it does not simulate a strong “gestalt switch” (Kuhn, 1962) between representations. While we have modeled revolutionary local changes to sets of explanations (Chapter 6) or representations (Chapter 8), the propagation and belief revision across contexts is an incremental, evolutionary process. This propagation process is more amenable to Toulmin’s (1972) model of conceptual change in science, which abandons a discrete notion of “before and after.”

#### **9.4 Future work and limitations**

Conceptual change is vast. In terms of time, psychological conceptual change in a domain such as force dynamics can take place over at least a decade (e.g., Ioannides and Vosniadou, 2002) and misconceptions are often retained despite years of formal instruction (Clement, 1982). In terms of information, human conceptual change is promoted by specialized curricula (e.g., Brown, 1994) and hindered by years of using productive misconceptions (Smith, diSessa, and Roschelle, 1993). In terms of cognitive processes, conceptual change is driven by model-based reasoning (Nersessian, 2007; Griffith et al., 2000), analogy (Brown & Clement, 1989; Gentner et al., 1997; Carey, 2009), anomaly (Chinn & Brewer, 1998; Posner et al., 1982), explanation construction (Chi et al., 1994a; Sherin et al., 2012), social factors (Pintrich, Marx, & Boyle, 1993), and belief refutation (Chi, 2000; 2008). There is much to be done to model the full range of this phenomenon.

Our model may be extended to capture more aspects of psychological conceptual change along all of these dimensions. Each opportunity for extension represents a current limitation in our model, so we discuss these in tandem. We also discuss how a model of conceptual change might be practically applied in other software systems.

#### **9.4.1 Simulating over larger timescales**

While the simulations presented here capture the qualitative characteristics and trajectories of psychological conceptual change, the changes occur over many orders of magnitude fewer stimuli than students. To capture a humanlike timescale of conceptual change, we need to adjust (1) the system's response to explanation failure and (2) the number and nature of training data.

Our model is more proactive than students in terms of changing its knowledge. One reason for this is that people have many responses to anomalous data besides revising their domain knowledge. Several anomaly-response actions have been identified in Chinn & Brewer (1993; 1998), such as ignoring anomalous data, holding the data in abeyance, exempting the data from a theory's applicability, and re-explaining the data to fit within a theory. Feltovich et al. (2001) identifies additional tactics people employ to prevent making changes to domain knowledge. Implementing additional strategies for explanation failure will slow the rate of conceptual change in simulation. Making these decisions requires access to metaknowledge about the to-be-changed beliefs, much of which is already available in the explanation-based network.<sup>45</sup>

Modeling conservatism in revising domain knowledge can help us understand the factors that

---

<sup>45</sup> Some relevant metaknowledge already included in the model: (1) the number of (preferred) explanations supported by a belief; (2) the ratio of preferred explanations to non-preferred explanations supported by a belief; (3) the alternate explanations for explanandums; (4) concept-level preferences between beliefs and model fragments; and (5) the conditional probability of using some belief in an explanation given another belief is also used.

make misconceptions resilient, and it might have a practical benefit of helping cognitive systems avoid unnecessary computation.

Another reason why conceptual change takes much longer in people is that people must sift the relevant from the irrelevant, and deal with incomplete and noisy information. The training data in Chapters 5 and 8 are automatically encoded from comic graphs, which we believe is an important first step in simulating conceptual change over larger timescales; even so, the stimuli are sparser than observations in the real world. All else being equal, adding extraneous entities and relations to the training data will make analogical retrieval less effective and delay the discovery of qualitative proportionalities via analogy (e.g., in Chapter 8), which will slow the rate of learning. Additionally, the comic graphs segment each observation into meaningful qualitative states, where the real world is continuous. Since the system derives quantity changes from these states rather than observing them directly, it does not have to differentiate quantities such as speed, velocity, and acceleration, which is difficult for novice students (Dykstra et al., 1992). Using a 3D physics engine as a learning environment (e.g., Mugan and Kuipers, 2010) is a promising direction for providing more realistic stimuli, though sparseness is still an issue.

Memory retrieval might also contribute to the duration of human conceptual change. In our model, changes are propagated by (1) encountering a new scenario that needs explaining, (2) retrieving previous, similar scenarios from memory and then (3) using the models and categories from the previous explanations to solve a new problem. Since people are most often reminded of literally similar phenomena (Forbus, Gentner, and Law, 1995), they might fail to reuse models and categories to explain entire classes of relevant – but not literally similar – a phenomena. This would produce tightly-contextualized mental models, as is evident in Collins and Gentner's (1987) study of novice mental models of evaporation and diSessa et al.'s (2004) study of novice

mental models of force. As a result, when there are more analogs in memory, mental models could become more tightly contextualized, and conceptual change might become more difficult.

A final consideration for the timescale of conceptual change is that presently, our simulations perform conceptual change in isolation. If the agent had other operations and incoming observations to attend to, then it could not dedicate as much time to retrospective explanation operations. This means that a greater share of its observations would be explained using outdated knowledge, all else being equal. This would ultimately increase the likelihood that outdated knowledge gets reused and propagated, delaying the rate of change.

#### **9.4.2 Improving explanation construction**

Our model considers more possibilities than people seem to consider when it constructs explanations. For instance, in one of the simulation trials in Chapter 7, the system constructs 16 distinct explanations for why Chicago's seasons change. It then evaluates each explanation and chooses the explanation that the student gives in the interview transcript. However, the corresponding student in the study seems to incrementally generate a single explanation for Chicago's seasons over several minutes.

One solution to this problem is to turn our abductive model formulation algorithm into an incremental beam search. This would mean that as it back-chains and instantiates model fragments, the algorithm only considers the lowest cost (i.e., simplest or most probable) alternative that it has not yet considered. This would construct a single explanation, but the problem of estimating which path is lowest cost is difficult without looking ahead. Another idea for focusing search is to use other explanations to guide the search for a new explanation: if other, preferred explanations tend to chain from model fragment A to model fragment B over

other alternatives, do the same in this case. Alternatively, the system could apply the old explanation via analogical mapping and inference. Since analogical inference is not necessarily sound, the system could perform focused abductive reasoning to fill in gaps in justification structure that are not inferred.

Another solution is to keep the same model formulation algorithm and implement a greedy walk through the resulting justification structure to only reify a single well-founded explanation at a time. This back-loads the work, since when the system needs to perform belief revision, it will have to consider alternative paths through justification structure.

### 9.4.3 Improving explanation evaluation

We have described two means of computing preferences between explanations: cost functions and rules. However, these are only as effective as the cost bases and the contents of the rules, respectively. At present, we do not believe that either of these is complete. One gap in our cost function is that it only penalizes for *inclusion* of artifacts such as contradictions and assumptions, but it does not penalize for *omission* of beliefs within an explanation. For example, a student might be confident that the tilt of the earth is related to the changing of the seasons, but unsure of the specific mechanics (e.g., Sherin et al., 2012). Consequently, any explanation the student constructs that omits the earth's tilt should be penalized. This might be simulated by encoding a metaknowledge relation to conceptually associate the belief that the earth has a tilted axis with the belief that the seasons change.

Rules and cost functions might also be extended to capture other psychological *explanatory virtues* (Lombrozo, 2011). For instance, we can compute the conditional probability of multiple inferences to determine an explanation's perceived probability. Other explanatory virtues

include the diversity of knowledge within an explanation, scope, fruitfulness, goal appeal, and fit within a narrative structure. Some of these may be computable based on the metaknowledge in the network structure, while others, such as narrative structure, might require comparison to other explanations and generalizations.

#### **9.4.4 Other types of agency**

Right now, our system explains quantity changing using knowledge of physical mechanisms, but physical mechanisms are only type of causality. Dennett (1987) and Keil & Lockhart (1999) identify three main types of causality: (1) mechanical, which we address here; (2) intentional; and (3) teleology/design/function. Most adults explain why a boat floats via mechanical causality, using knowledge of density and buoyancy. Piaget (1930) found that children frequently ascribe intentions (e.g., the boat doesn't want to sink) or teleology (e.g., it floats so we can ride on top of the water) to physical situations. This results in anthropocentric finalism, where natural phenomena are explained relative to their function for humans, or animism, where nonliving things are assigned lives and intentions. Having the system learn when to use which agency, e.g., by contextualizing and reusing them by similarity or by using modeling assumptions (discussed below), is an interesting opportunity to model these aspects of cognitive development as conceptual changes.

For example, it is possible that the two students in Chapter 7 who were not modeled by our simulation arrived at their final model using teleological explanation. Recall that the two students who were not simulated in Chapter 7 were in the “prompted” condition, where students explained to themselves while reading. Both used the incorrect “single loop (lung)” model of the circulatory system at the posttest, where blood flows from the heart, through the lungs, to the

body, and back. These students generated erroneous components within their mental models through self-explanation. More specifically, they might have (1) understood that the function of the lungs is to oxygenate the blood for eventual delivery to the body and (2) inferred the structure of the circulatory system by attending to this lung function.

#### 9.4.5 Taking analogy further

In Chi et al. (1994), students made spontaneous analogies such as “the septum [of the heart] is like a wall that divides the two parts” when explaining textbook material. While our system uses analogy to retrieve similar examples and infer qualitative proportionalities (Chapter 8), it does not make spontaneous analogies to transfer knowledge across domains. Analogical inference is a powerful strategy worth incorporating into our model of explanation construction. We can sketch this idea very generally. As new information (e.g., about the septum dividing the sides of the heart) is incorporated via reading, it can be used as a probe to MAC/FAC, which can retrieve similar concepts (e.g., a wall dividing two spaces). The SME mapping between the new and existing concepts produces candidate inferences, which can elaborate the new material with respect to surface characteristics, function, and causal structure. As mentioned in Chapter 3, analogical inferences might not be deductively valid, so this might produce additional misconceptions (Spiro et al 1989).

When analogies are communicated through instruction or text, they have the capability to foster conceptual change (Brown, 1994; Gentner et al., 1997; Vosniadou et al., 2007). These are important considerations for extending the system further. For example, *bridging analogies* (Brown & Clement, 1989) can be used to facilitate the transfer of knowledge from a correct base scenario (e.g., an outstretched hand exerts an upward force on a book at rest on its surface) to a



flawed target scenario (e.g., a table *does not* exert an upward force on a book at rest on its surface). Through a sequence of bridging analogies, such as a book on a spring, a book on a mattress, and a book on a pliable board, beliefs are imported into the target scenario. This permits the construction of new explanations that can replace the old, flawed explanations. Since analogical mapping and transfer are built into Companions cognitive architecture, this is a reasonable next step.

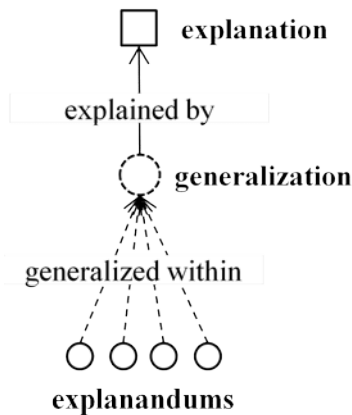
#### **9.4.6 Accruing domain knowledge**

The simulations in Chapters 5 and 8 acquire model fragments by induction and heuristics, respectively. By contrast, the simulations in Chapters 6 and 7 start with hand-coded model fragments, based on pretests, posttests, and interview transcripts with students. In these simulations, we did not model how the initial qualitative models of contained fluid, fluid flow, fluid enrichment, astronomical heating, astronomical orbit, and so-forth, are acquired by the students. Presumably, people learn about these processes and relationships by some combination of interaction, reading, and observation, and hand-coding these representations is not good practice for cognitive modeling in the long term. A more ideal solution is to automatically encode the initial knowledge of a student using a natural language understanding (NLU) system with deep semantic interpretation (e.g., Tomai & Forbus, 2009) to analyze an interview transcript in order to automatically construct the initial set of model fragments.

Acquiring new qualitative model fragments from text is an unsolved problem, but there have been advances in deriving QP theory interpretations from natural language (Kuehne, 2004), which is an important component of learning reusable models.

### 9.4.7 Storing explanations

Preferred explanations are central organizing structures in our model, so they persist over time. Explanations that are *not* preferred also persist over time in our model because they might eventually become preferred through a belief revision process, as in Chapter 7. Explanations are very compact<sup>46</sup> in our system, but the justification structure requires considerably more storage. We did not encounter a performance degradation or storage bottleneck due to the algorithms and explanation-based knowledge organization described here, but problems could arise if we imagine a lifetime of experience and learning. These are important considerations for cognitive modeling as well as for performance over time.



**Figure 55: Using SAGE to cluster explanandums so that one explanation can justify multiple observations that are close analogs of one another.**

Storing the justification structure for all of the explanations in our system saves computation, but it creates a potential storage bottleneck. We could feasibly store each explanation as  $\langle B, M \rangle$ , and re-derive the justification structure when necessary using the explanation construction algorithm over the beliefs and model fragments in  $B$  alone. This would constrain the search for explanations to only the beliefs and model fragments within the previous

<sup>46</sup> Each explanation  $\langle J, B, M \rangle$  is lightweight because the set of beliefs  $B$  and explanandums  $M$  are determinable based on the set of justifications  $J$ . Consequently, the storage requirement for each explanation includes a symbol for itself and a set of symbols indicating its justifications.

explanation. This relaxes the psychological assumption that people retain all of the justifications for their beliefs, but it still assumes that people retain their preferred explanations.

#### 9.4.8 Clustering explanandums

In addition to retaining its preferred explanations, the system retains its preferred explanation *for each explanandum*. This means that whenever a new phenomenon is explained, a preferred explanation is associated with that exact phenomenon in memory. As explanandums are encountered and explained, this may become intractable, so this it might be an unrealistic psychological assumption. One way to relax this assumption is to (1) use analogical generalization to cluster explanandums using unsupervised learning and then (2) explain each generalization. This is illustrated in Figure 55.

This saves space as well as computation. For instance, consider that the agent must explain why a ball rolls to the left after being kicked, and it has a SAGE generalization describing examples of this very phenomenon. If the generalization has already been explained by some explanation  $x$ , then no first-principles reasoning has to occur to explain the ball rolling to the left – the agent merely has to construct an analogical mapping to the generalization and ensure that the generalization’s explanation  $x$  holds on the new explanandum. This means that the system would only generate new explanations if it encounters an explanandum that is not structurally similar to a previous generalization or explanandum.

This idea of generalized explanandums is similar to the idea of storing *prototype histories* (Forbus & Gentner, 1986) which describe generalizations of phenomena occurring over time. We have demonstrated in (Friedman and Forbus, 2008) that SAGE can learn these from examples, so it is a reasonable optimization.

#### 9.4.9 Proactivity

These simulations perform conceptual change as a result of observing the world and receiving instructional material. This does not capture the more active aspects of human learning, such as asking questions, planning, experimenting, and teaching others. User interaction and user modeling are central goals of the Companions cognitive architecture (Forbus et al., 2009) within which this model is implemented, so progress is being made on several of these social interaction fronts. In terms of experimentation, the present model provides some support for active learning. For example, provided the hypothesis *the distance a box slides is inversely proportional to its size*, the system might test this hypothesis by retrieving previous example, increasing the size of the object, and requesting a training datum of the modified observation. Provided this new, solicited observation, the system could detect and resolve explanation failures as already described in Chapter 8.

#### 9.4.10 Applying the model of conceptual change

The model of conceptual change presented here might be practically applied within intelligent tutoring systems (ITS; e.g. Koedinger et al., 1997). ITSs automatically deliver customized feedback to a student based on the student's performance. They often include a *task model* to represent expert knowledge and a *student model* to track student knowledge. Both are crucial for diagnosing student misconceptions, tracking progress, and selecting new problems to maximize learning. Our computational model of conceptual change uses a single knowledge representation strategy to represent student misconceptions and correct scientific concepts in several scientific domains, including dynamics, biology, and astronomy. Consequently, the model might ultimately be integrated into ITSs to (1) represent an extendable library of student knowledge,

(2) discover a student's mental models using active learning in a tutoring session, (3) find inconsistencies in a student's mental models via abduction and qualitative simulation, and (4) guide the student through a curriculum to confront and remedy the inconsistencies according to a simulation of conceptual change using his or her mental models. This requires substantial additional work, but progress in using compositional modeling for tutoring systems has been made by de Koning et al. (2000) and others. It could lead to Socratic tutors that have human-like flexibility (e.g., Stevens & Collins, 1977).

## References

- Abdelbar, A. M. (2004). Approximating cost-based abduction is NP-hard. *Artificial Intelligence*, 159: 231-239.
- Alchourròn, C. E., Gärdenfors, P., and Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50: 510–530.
- Amir, E. and McIlraith, S. (2005). Partition-Based Logical Reasoning for First-Order and Propositional Theories. *Artificial Intelligence*, 162(1-2): 49-88
- Anderson, J. R., and Gluck, K. (2001). What role do cognitive architectures play in intelligent tutoring systems. In D. Klahr & S. M. Carver (Eds.) *Cognition and Instruction: Twenty-five Years of Progress*: 227-262. Mahwah, NJ: Erlbaum.
- Atwood, R. K., and Atwood, V. A. (1996). Preservice elementary teachers' conceptions of the causes of seasons. *Journal of Research in Science Teaching*, 33(5), 553–563.
- Baillargeon, R. (1998). A Model of Physical Reasoning in Infancy. In Lipsitt, L. P. and Rovee-Collier, C. (Eds.), *Advances in infancy research*, 9.
- Bridewell, W. & Langley, P. (2011). A Computational Account of Everyday Abductive Inference. *Proceedings of the 33<sup>rd</sup> Annual Meeting of the Cognitive Science Society*. Boston, MA.
- Brown, D. E. and Clement, J. (1989) Overcoming misconceptions by analogical reasoning: abstract transfer versus explanatory model construction. *Instructional Science*, 18: 237-261.
- Brown, D. (1994). Facilitating conceptual change using analogies and explanatory models. *International Journal of Science Education*, 16(2), 201-214.
- Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. and Mitchell, T. (2010). Toward an architecture for Never-Ending Language Learning. *Proceedings of AAAI-10*.

- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Carey, S. (1988). Reorganization of knowledge in the course of acquisition. In: Sidney Strauss, Ed. *Ontogeny, phylogeny, and historical development*. 1-27.
- Carey, S. (1991). Knowledge acquisition-enrichment or conceptual change? In S. Carey, & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 257-292). Hillsdale, NJ: Erlbaum.
- Carey, S. (2009). *The Origin of Concepts*. New York, NY: Oxford University Press.
- Cassimatis, N. (2006). A cognitive substrate for achieving human-level intelligence. *AI Magazine*. 27(2), 45-56.
- Cassimatis, N., Bello, P., and Langley, P. (2008). Ability, Breadth and Parsimony in Computational Models of Higher-Order Cognition. *Cognitive Science*, 32(8): 1304-1322.
- Charniak, E. and Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. In *Proceedings of AAAI National Conference on Artificial Intelligence*: 446-451.
- Charniak, E. and Shimony, S. E. (1994). Cost-based abduction and MAP explanation. *Artificial Intelligence*, 66: 345-374.
- Chen, Z. (1995). Analogical transfer: From schematic pictures to problem solving. *Memory & Cognition*, 23(2): 255-269.
- Chen, Z. (2002). Analogical Problem Solving: A Hierarchical Analysis of Procedural Similarity. *Journal of Experimental Psychology: Learning Memory, and Cognition*, 28(1): 81-98.
- Chi, M., de Leeuw, N., Chiu, M., and LaVancher, C. (1994a). Eliciting self-explanations improves understanding. *Cognitive Science*, 18: 439-477.
- Chi, M., Slotta, J. D. and de Leeuw, N. (1994b). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction*, 4: 27-43.

- Chi, M. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology*, Hillsdale, NJ: Lawrence Erlbaum Associates. 161-238.
- Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25: 471-533.
- Chi, M. (2005). Commonsense Conceptions of Emergent Processes: Why Some Misconceptions Are Robust. *The Journal of the Learning Sciences*, 14(2): 161-199.
- Chi, M. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *Handbook of research on conceptual change* (pp. 61-82). Hillsdale, NJ: Erlbaum.
- Chi, M. T. H. and Brem, S. K. (2009). Contrasting Ohlsson's resubsumption theory with Chi's categorizational shift theory. *Educational Psychologist*, 44(1): 58-63.
- Chinn, C. and Brewer, W. (1993). The role of anomalous data in knowledge acquisition: a theoretical framework and implications for science instruction. *Review of Educational Research*, 63(1): 1-49.
- Chinn, C. and Brewer, W. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, 35(6): 623-654.
- Chinn, C., Brewer, W., and Samarapungavan, A. (1998). Explanation in Scientists and Children. *Minds and Machines*, 8(1): 119-136.
- Christie, S. and Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, 11(3): 356-373.
- Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50(1): 66-71.



- Collins, A. and Gentner, D. (1987). How people construct mental models. In D. Holland & N. Quinn (Eds.), *Cultural models in language and thought* (pp. 243-265). England: Cambridge University Press.
- Cox, M. (2005). Metacognition in Computation: A selected research review. *Artificial Intelligence*, 169(2): 104-141.
- Cox, M. T. and Raja, A. (2007). Metareasoning: A Manifesto. *Technical Report BBN TM 2028*, BBN Technologies.
- Cox, M. T. and Ram, A. (1999). Introspective Multistrategy Learning: On the Construction of Learning Strategies. *Artificial Intelligence*, 112: 1-55.
- Davies, M. F. (1997). Belief Persistence after Evidential Discrediting: The Impact of Generated versus Provided Explanations on the Likelihood of Discredited Outcomes. *Journal of Experimental Social Psychology*, 33(6): 561-578.
- de Kleer, J., and Brown, J. S. (1984). A qualitative physics based on confluences. *Artificial Intelligence*, 24(1-3): 7-83.
- de Koning, K., Bredweg, B., Breuker, J., and Wielinga, B. (2000). Model-based reasoning about learner behavior. *Artificial Intelligence*, 117(2): 173-229.
- de Leeuw, N. (1993). Students' beliefs about the circulatory system: Are misconceptions universal? *Proceedings of the 15<sup>th</sup> Annual Meeting of the Cognitive Science Society*.
- de Leeuw, N. & Chi, M.T.H. (2003). The role of self-explanation in conceptual change learning. In G. Sinatra & P. Pintrich (Eds.), *Intentional Conceptual Change*. Erlbaum: 55-78.
- DeJong, G. (Ed.) (1993). *Investigating Explanation-Based Learning*. Kluwer Academic Publishers, Norwell, MA, USA.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

- diSessa, A. (1988). Knowledge in pieces. In G. Foreman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49-70). Mahwah, NJ: Lawrence Erlbaum Associates.
- diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction*, 10(2-3): 105-225.
- diSessa, A., Gillespie, N., Esterly, J. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28, 843-900.
- diSessa, A. A. (2006). A history of conceptual change research: threads and fault lines. In K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences*: 265-282. MA: Cambridge University Press.
- Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, 12(3): 231-272.
- Doyle, J. and Wellman, M. (1990). Rational distributed reason maintenance for planning and replanning of large-scale activities. *DARPA Workshop on Innovative Approaches to Planning, Scheduling, and Control*.
- Doyle, J. (1991). Rational Belief Revision. *Proceedings of the Second Conference on Principles of Knowledge Representation and Reasoning*: 163-174.
- Doyle, J. (1992). Reason maintenance and belief revision. In P. Gärdenfors (Ed.), *Belief Revision*, Cambridge: Cambridge University Press. 29-51.
- Dykstra, D. I., Boyle, C. F., & Monarch, I. A. (1992). Studying conceptual change in learning physics. *Science Education*, 76(6): 615-652.
- Esposito, F., Semeraro, G., Fanizzi, N., & Ferilli, S. (2000). Conceptual Change in Learning Naive Physics: The Computational Model as a Theory Revision Process. In E. Lamma and P. Mello (Eds.), *AI\*IA99: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence 1792*, 214-225, Springer: Berlin.

- Esposito, F., Semeraro, G., Fanizzi, N. & Ferilli, S. (2000). Multistrategy Theory Revision: Induction and Abduction in INTHELEX. *Machine Learning Journal*, 38(1,2): 133-156.
- Falkenhainer, B., Forbus, K. and Gentner, D. (1989). The Structure Mapping Engine: Algorithm and examples. *Artificial Intelligence*, 41: 1-63.
- Falkenhainer, B. & Forbus, K. (1991). Compositional modeling: Finding the right model for the job. *Artificial Intelligence*, 51: 95-143.
- Faries, J. M., & Reiser, B. J. (1988). Access and use of previous solutions in a problem-solving situation. *Proceedings of the tenth annual meeting of the Cognitive Science Society*.
- Feltovich, P., Coulson, R., and Spiro, R. (2001). Learners' (mis) understanding of important and difficult concepts. In K. Forbus & P. Feltovich, (Eds.) *Smart Machines for Education*. 349-375.
- Forbus, K. (1992). Pushing the edge of the (QP) envelope. In B. Faltings and P. Struss, (Eds.) *Recent Progress in Qualitative Physics*. MIT Press.
- Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, 24: 85-168.
- Forbus, K. and Gentner, D. (1986). Learning Physical Domains: Towards a Theoretical Framework. In Michalski, R., Carbonell, J. and Mitchell, T. (Eds.), *Machine Learning: An Artificial Intelligence Approach, Volume 2*. Tioga press.
- Forbus, K. and de Kleer, J. (1993). *Building Problem Solvers*, MIT Press.
- Forbus, K., Klenk, M., and Hinrichs, T. (2009). Companion Cognitive Systems: Design Goals and Lessons Learned So Far. *IEEE Intelligent Systems*, 24(4): 36-46.
- Forbus, K. D., Usher, J., Lovett, A., Lockwood, K., and Wetzel, J. (2008). CogSketch: Open-domain sketch understanding for cognitive science research and for education. *Proceedings*

*of the Fifth Eurographics Workshop on Sketch-Based Interfaces and Modeling*. Annecy, France.

Forbus, K. D., Gentner, D. and Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19(2): 141-205.

Friedman, S. and Forbus, K. (2008). Learning Qualitative Causal Models via Generalization & Quantity Analysis. *Proceedings of the 22nd International Workshop on Qualitative Reasoning*. Boulder, CO.

Friedman, S., Taylor, J., and Forbus, K. (2009). Learning Naive Physics Models by Analogical Generalization. *Proceedings of the 2nd International Analogy Conference*. Sofia, Bulgaria.

Friedman, S. and Forbus, K. (2010). An Integrated Systems Approach to Explanation-Based Conceptual Change. *Proceedings of AAAI-10*. Atlanta, GA.

Friedman, S. and Forbus, K. (2011). Repairing Incorrect Knowledge with Model Formulation and Metareasoning. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. Barcelona, Spain.

Friedman, S., Forbus, K., and Sherin, B. (2011). How do the seasons change? Creating & revising explanations via model formulation & metareasoning. *Proceedings of the 25th International Workshop on Qualitative Reasoning*. Barcelona, Spain.

Friedman, S., Lovett, A., Lockwood, K., McLure, M., and Forbus, K. (in prep). SAGE: A model of analogical generalization.

Gärdenfors, P. (1990). The dynamics of belief systems: Foundations vs. coherence theories. *Revue Internationale de Philosophie*, 172: 24-46.

Gentner, D. and Stevens, A. L. (1983). *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum.

- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2): 155-170.
- Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199-241). London: Cambridge University Press. (Reprinted in *Knowledge acquisition and learning*, 1993, 673-694).
- Gentner, D., Rattermann, M. J., and Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25: 524-575.
- Gentner, D. and Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, 14: 487-513.
- Gentner, D. (2002). Mental models, psychology of. In N. J. Smelser & P. B. Bates (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 9683-9687). Amsterdam: Elsevier Science.
- Gentner, D., Levine, S., Dhillon, S. and Poltermann, A. (2009). Using structural alignment to facilitate learning of spatial concepts in an informal setting. *Proceedings of the Second International Conference on Analogy*. Sofia, Bulgaria.
- Gentner, D., Loewenstein, J., Thompson, L., and Forbus, K. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, 33: 1343-1382.
- Gick, M. L. and Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12(3): 306-355.
- Glenberg, A. M., Robertson, D. A., Jansen, J. L., and Johnson-Glenberg, M. C. (1999). Not propositions. *Journal of Cognitive Systems Research*, 1: 19-33.

- Griffith, T. W., Nersessian, N. J., and Goel, A. (1996). The role of generic models in conceptual change. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*.
- Griffith, T. W., Nersessian, N. J., and Goel, A. (2000). Function-follows-form transformations in scientific problem solving. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*.
- Halloun, I., and Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, 53(1043).
- Hempel, C.G. and Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, XV: 135–175.
- Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30: 141-158.
- Horvitz, E. (1988). Reasoning about beliefs and actions under computational resource constraints. *Uncertainty in Artificial Intelligence*, 3: 301–324.
- Inagaki, K. and Hatano, G. (2002). *Young children's naïve thinking about the biological world*. New York: Psychology Press.
- Ioannides, C., and Vosniadou, S. (2002). The changing meanings of force. *Cognitive Science Quarterly*, 2: 5-61.
- Kass, A. (1994). TWEAKER: Adapting old explanations to new situations. In R. Schank, A. Kass, and C. K. Riesbeck (Eds.), *Inside Case-Based Explanation*, New Jersey: Lawrence Erlbaum Associates, Inc.
- Katsuno, H., and Mendelzon, A. (1991). Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52: 263-294.

- Kennedy, C. M. (2008). Distributed Meta-Management for Self-Protection and Self-Explanation. In *Proceedings of the AAAI-08 Workshop on Metareasoning*.
- Keil, F. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. Hirschfield & S. Gelman (eds.) *Mapping the mind: Domain specificity in cognition and culture*: 234-254. Cambridge, UK: Cambridge University Press.
- Keil, F. C. and Lockhart, K. L. (1999). Explanatory understanding on conceptual development. In Scholnick, E. K. ed. *Conceptual development: Piaget's legacy*.
- Koedinger, K. R., Anderson, J. R., Hadley, W. H., and Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1): 30-43.
- Kotovskiy, L. and Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67: 2797-2822.
- Kuehne, S. E. (2004). *Understanding natural language descriptions of physical phenomena* (Tech. Report NWU-CS-04-32). Doctoral dissertation, Northwestern University, Evanston, Illinois.
- Kuipers, B. (1986). Qualitative simulation. *Artificial Intelligence*, 29(3): 289-338.
- Kuhn, T. (1962). *The Structure of Scientific Revolutions*. The University of Chicago.
- Laird, J., Newell, A., and Rosenbloom, P. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1): 1-64.
- Langley, P. (2000). The computational support of scientific discovery. *International Journal of Human-Computer Studies*, 53: 393-410.

- Langley, P., Bradshaw, G. L., and Simon, H. A. (1983). Rediscovering chemistry with the Bacon system. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. San Mateo, CA: Morgan Kaufmann.
- Leake, D. and Wilson, M. (2008). Extending Introspective Learning from Self-Models. In *Proceedings of the AAAI-08 Workshop on Metareasoning*.
- Leake, D. (1992). *Explanations: A Content Theory*. New Jersey: Lawrence Erlbaum Associates, Inc.
- Lenat, D. (1983). EURISKO: A program that learns new heuristics and domain concepts. *Artificial Intelligence*, 21: 61-98.
- Lenat, D. B., and Brown, J. S. (1984). Why AM and EURISKO appear to work. *Artificial Intelligence*, 23(3): 269-294.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10): 464-470.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55: 232-257.
- Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, 6/8: 539-551.
- Markman, A. B. and Gentner, D. (2001). Thinking. *Annual Review of Psychology*, 52: 223-247.
- Marr, D. (1982). *Vision: A Computational Approach*. San Francisco: Freeman & Co.
- McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence*, 171(1): 1174-1182.
- McCloskey, M. (1983). Naïve theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum. 299-323.



- McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*, 57: 4-9.
- McLure, M., Friedman, S., and Forbus, K. (2010). Learning concepts from sketches via analogical generalization and near-misses. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society (CogSci)*. Portland, OR.
- Minstrell, J. (1982). Explaining the “at rest” condition of an object. *The Physics Teacher*, 20: 10-14.
- Minstrell, J. (1989). Teaching science for understanding. In L. Resnick, & L. Klopfer (Eds.), *Toward the thinking curriculum* (pp. 129-149). Alexandria, VA: Association for Supervision and Curriculum Development.
- Minton, S. (1990). Quantitative results concerning the utility of explanation-based learning. *Artificial Intelligence*, 42(2-3).
- Molineaux, M., Kuter, U., and Klenk, M. (2011). What just happened? Explaining the past in planning and execution. *Proceedings of the Sixth International Workshop on Explanation-aware Computing*.
- Mugan, J. and Kuipers, B. (2011). Autonomous Learning of High-Level States and Actions in Continuous Environments. *IEEE Transactions on Autonomous Mental Development*.
- Nersessian, N. J. (2007). Mental Modeling in Conceptual Change. In S. Vosniadou (Ed.) *International Handbook of Conceptual Change*. London: Routledge. 391-416.
- Ng, H. T. and Mooney, R. J. (1992). Abductive plan recognition and diagnosis: A comprehensive empirical evaluation. *Proceedings of the 3<sup>rd</sup> International Conference on Principles of Knowledge Representation & Reasoning*: 499–508.
- Nilsson, N. (2005). Human-level Artificial Intelligence? Be Serious! *AI Magazine*.

- Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14: 510-520.
- Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist*, 44(1): 20–40.
- Paritosh, P.K. (2004). Symbolizing Quantity. In *Proceedings of the 26<sup>th</sup> Annual Meeting of the Cognitive Science Society*.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.
- Peirce, C. S. (1958). *Collected Papers of Charles Sanders Peirce*. Cambridge, Mass.: MIT Press.
- Pintrich, P. R., Marx, R. W., and Boyle, R. A. (1993). Beyond cold conceptual change: the role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, 63(2): 167-199.
- Poole, D. L. (1985). On the comparison of theories: preferring the most specific explanation. In *Proceedings of IJCAI-85*: 144-147.
- Posner, G. J., Strike, K. A., Hewson, P. W., and Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2): 211-227.
- Quine, W. V. O. and Ullian, J. S. (1970). *The Web of Belief*. New York: Random House.
- Quine, W. V. O. (1960). *Word and Object*. Oxford, England: MIT Press.
- Raghavan, S., and Mooney, R. (2010). Bayesain Abductive Logic Programs. *Proceedings of the AAAI-10 Workshop on Statistical Relational AI*: 81-87.

- Ram, A. and Cox, M. (1994). Introspective reasoning using meta-explanations for multistrategy learning. In Michalski, R., and Tecuci, G., eds., *Machine Learning: A Multistrategy Approach*. Morgan Kaufmann. 349-377.
- Reif, F. (1985). Acquiring an effective understanding of scientific concepts. In L. H. T. West & A. L. Pines (Eds.) *Cognitive Structure and Conceptual Change*, Academic Press, Inc.
- Reiner, M., Slotta, J. D., Chi, M. T. H., and Resnick, L. B. (2000). Naïve Physics Reasoning: A Commitment to Substance-Based Conceptions. *Cognition and Instruction*, 18(1): 1-34.
- Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62: 107-136.
- Rickel, J. and Porter, B. (1995) Automated modeling of complex systems to answer prediction questions. *Artificial Intelligence*, 93(1-2), 201-260.
- Rips, L. J. (1986). Mental Muddles. In M. Brand and R. M. Harnish (eds.), *The Representation of Knowledge and Belief*. Tucson, AZ: University of Arizona Press. 258-286.
- Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, 12: 129-134.
- Ross, L. and Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, and A. Tversky, eds. *Judgement under Certainty: Heuristics and Biases*. Cambridge: Cambridge University Press. 129-152.
- Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13: 629-639.
- Rottman, B. M. and Keil, F. C. (2011). What matters in scientific explanations: Effects of elaboration and content. *Cognition*.

- Russell, S. J. and Wefald, E. (1991). Principles of Metareasoning. *Artificial Intelligence*, 49: 361-395.
- Santos, E. (1994). A linear constraint satisfaction approach to cost-based abduction. *Artificial Intelligence*, 65: 1-27.
- Schank, R. C. (1986). *Explanation patterns: understanding mechanically and creatively*. New Jersey: Lawrence Erlbaum Associates, Inc.
- Schank, R. C., Kass, A., and Riesbeck, C. K. (1994). *Inside Case-based Explanation*, New Jersey: Lawrence Erlbaum Associates, Inc.
- Sherin, B. (2006). Common sense clarified: Intuitive knowledge and its role in physics expertise. *Journal of Research in Science Teaching*: 33(6), 535-555.
- Sherin, B. L., Krakowski, M., and Lee, V. R. (2012). Some assembly required: how scientific explanations are constructed during clinical interviews. *Journal of Research in Science Teaching*, 49(2): 166-198.
- Sinatra, G. M. and Pintrich, P. R. (2003). *Intentional Conceptual Change*. Lawrence Erlbaum Associates, Inc.
- Singla, P. and Mooney, R. (2011). Abductive Markove Logic for Plan Recognition. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*: 1069-1075.
- Smith, J. P., diSessa, A. A., and Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3(2).
- Spiro, R., Feltovich, P., Coulson, R., and Anderson, D. (1989). Multiple analogies for complex concepts: antidotes for analogy-induced misconception in advanced knowledge acquisition. In Vosniadou, S. and Ortony, A. (Eds.) *Similarity and Analogical Reasoning*. pp. 498-531, Cambridge University Press.

- Stevens, A. L. and Collins, A. (1977). The goal structure of a Socratic tutor. In *Proceedings of the Association for Computing Machinery Annual Conference*. Association for Computing Machinery.
- Taylor, J. L. M., Friedman, S. E., Forbus, K. D., Goldwater, M., and Gentner, D. (2011). Modeling structural priming in sentence production via analogical processes. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Boston, MA.
- Thagard, P. (2000). Probabilistic Networks and Explanatory Coherence. *Cognitive Science Quarterly*, 1: 93-116.
- Thelen, E. and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: The MIT Press
- Timberghien, A. (1994). Modeling as a basis for analyzing teaching-learning situations. *Learning and Instruction*, 4: 71-87.
- Tomai, E. and Forbus, K. (2009). EA NLU: Practical Language Understanding for Cognitive Modeling. *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference*.
- Toulmin, S. (1972). *Human understanding*. Vol. 1. Oxford, UK: Clarendon Press.
- Trickett, S. B. and Trafton, G. (2007). “What If...”: The Use of Conceptual Simulations in Scientific Reasoning. *Cognitive Science*, 31: 843-875.
- Vosniadou, S. (1994). Capturing and Modeling the Process of Conceptual Change. In S. Vosniadou (Guest Editor), *Special Issue on Conceptual Change, Learning and Instruction*, 4: 45-69.

- Vosniadou, S. (2002). Mental Models in Conceptual Development. In L. Magnani & N. Nersessian (Eds.) *Model-Based Reasoning: Science, Technology, Values*. New York: Kluwer Academic Press.
- Vosniadou, S. (2007). The conceptual change approach and its re-framing. In S. Vosniadou, A., Baltas, , & X., Vamvakoussi, (Eds.) *Reframing the conceptual change approach in learning and instruction*. Oxford: Elsevier.
- Vosniadou, S., Skopeliti, I., and Gerakaki, S. L. (2007). Understanding the role of analogies in restructuring pro-cesses. In A. Schwering, U. Krumnack, K.U. Kuhnberger, & H. Gust (Eds.), *Analogies: Integrating Multiple Cognitive Abilities, Publications of the Institute of Cognitive Science vol. 5*: 39-43. Osnabruck, Germany.
- Vosniadou, S. and Brewer, W.F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24: 535-585.
- Vosniadou, S. and Brewer, W. F. (1994). Mental Models of the Day/Night Cycle. *Cognitive Science*, 18: 123-183.
- Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Education Research*, 13: 21-3.
- Winston, P. and Rao, S. (1990). Repairing learned knowledge using experience. *Massachusetts Institute of Technology, AI Memo #1231, May 1990*.
- Wiser, M. and Carey, S. (1983). When heat and temperature were one. In D. Gentner & A. L. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum: 267-297.
- Wiser, M. and Amin, T. (2001). “Is heat hot?” Inducing conceptual change by integrating everyday and scientific perspectives on thermal phenomena. *Learning and Instruction*, 11(4-5): 331-355.

## Appendix

### Definitions

We define several terms in the below table for ease of reference and clarity. Since we are concerned with learning over time, we use the term “memory” to refer to long-term memory, unless otherwise specified.

Term/Symbol	Definition
<i>belief</i>	A proposition represented as a relation <i>reln</i> and at least one arguments $\{a_0, \dots, a_n\}$ , written as $(reln\ a_0 \dots a_n)$ .
<i>model fragment belief</i>	A belief referring to the existence of a model fragment <i>m</i> , of the form $(isa\ m\ ModelFragment)$
<i>scenario microtheory</i>	A microtheory that includes beliefs and model fragments. Each scenario microtheory represents information gathered via observation, instruction, or other type of interaction.
$\mathbb{D} = \{b_0, \dots, b_n\}$	The <i>domain knowledge microtheory</i> , which contains beliefs that can be believed regardless of whether they are used in an explanation. This includes explanandums, model fragment beliefs, and other beliefs from observation and instruction.

	Inherits all beliefs from all scenario microtheories.
$\mathbb{D}_a = \{b_0, \dots, b_m\} \subseteq \mathbb{D}$	The <i>adopted domain knowledge microtheory</i> is the subset of the domain knowledge microtheory that is presently believed by the agent. For example, the agent may have the propositional belief “the heart oxygenates the blood” in $\mathbb{D}$ and not in $\mathbb{D}_a$ . This permits the agent to reason about a belief’s consequences without believing it.
<i>explanandum</i>	A phenomenon that requires an explanation, <sup>47</sup> represented as a set $m$ of one or more beliefs $\{b_0, \dots, b_m\}$ . In our simulations, these range from sets of multiple beliefs (e.g., describing flood flowing from the heart to the body in Chapter 7) or sets of single beliefs (e.g., describing quantity changes in Chapter 8). Each explanandum is believed in $\mathbb{D}$ .
$\mathbb{M} = \{M_0, \dots, M_n\}$	The set of all explanandums in the agent’s memory.
$\mathbb{B} = \{b_0, \dots, b_n\}$	The <i>provisional belief microtheory</i> , containing beliefs that are either assumed or inferred from other knowledge. Beliefs in this microtheory are only believed if they are in an explanation that is adopted by the agent.
<i>justification</i>	Rationale for belief. Includes rule-based inferences, model

<sup>47</sup> The term “explanandum” has been used to describe a phenomenon requiring an explanation. The explanandum is typically the subject of a *why* question rather than a *what* question (Hempel & Oppenheim, 1948).



---

	<p>fragment instantiations, model fragment activations, and other rationale. Each justification <math>j</math> has a set of <i>antecedent</i> beliefs <b><i>antes</i></b>(<math>j</math>) and a set of <i>consequence</i> beliefs <b><i>conseqs</i></b>(<math>j</math>), such that the conjunction of <b><i>antes</i></b>(<math>j</math>) is sufficient to believe <b><i>conseqs</i></b>(<math>j</math>).</p>
<i>explanation</i>	<p>Uniquely defined as <math>\langle J, B, M \rangle</math>. Represents a well-founded explanation <math>J \subseteq \mathbb{J}</math> for some explanandum(s) <math>M \subseteq \mathbb{M}</math>. <math>B</math> is a set of beliefs comprised of (1) beliefs supporting <math>M</math> through <math>J</math> and (2) metaknowledge<sup>48</sup> <math>B_m</math> about the explanation. More formally:</p> $B = B_m \cup \bigcup_{j \in J} \text{antes}(j) \cup \text{conseqs}(j)$
<i>explanation microtheory</i>	<p>A single explanation microtheory exists for each explanation <math>\langle J, B, M \rangle</math>. Contains all beliefs <math>B</math> in the explanation, and is a proper subset of one or more beliefs in <math>\mathbb{B}</math> and <math>\mathbb{D}</math>.</p>
<i>assumption</i>	<p>An unjustified belief <math>b \in \mathbb{B}</math> that is not part of the domain theory <math>\mathbb{D}</math>. More formally, <math>b</math> is assumed in an explanation <math>\langle J, B, M \rangle</math> if and only if it is part of the well-founded explanation but it is not itself justified.</p>
<i>explanation competition</i>	<p>Occurs between two different explanations <math>\langle J, B, M \rangle</math> and <math>\langle J', B', M' \rangle</math> for explanandum <math>m</math> if and only if <math>m \in (M \cap M')</math>.</p>

---

<sup>48</sup> In our simulations, metaknowledge about an explanation include the beliefs about the structure of an explanation, such as the presence of an asymmetric quantity change in a cyclic state space (e.g., Chapter 6). These beliefs affect how preferences are computed between explanations, but they do not affect the justification structure.

---

<i>preferred explanation</i>	For an explanandum $m$ , the agent's preferred explanation.
------------------------------	---

---

$\mathbb{E} = \{\langle m_0, x_0 \rangle, \dots, \langle m_n, x_n \rangle\}$	The <i>explanandum mapping</i> over every explanandum $m \in \mathbb{M}$ to its respective best explanation $x_i$ . Exhaustive over all explanandums $\mathbb{M}$ , but not necessarily over all explanations.
--	--

---

### Transcript of an interview about the seasons from Chapter 6

Below is a transcript of the student “Angela,” courtesy of Sherin et al. (2012). We have removed symbols that indicate gestures, emphasis, and pauses, but we have kept some nonverbal annotations where helpful for understanding the conversation. A = Student, B = Interviewer.

---

	Who	Transcript
--	-----	------------

---

1	B	I want to know why it's warmer in the summer and colder in the winter
2	A	That's because like the sun is in the center and the Earth moves around the sun and the Earth is at one point like in the winter it's like farther away from the sun-
3	B	uh huh-
4	A	and towards the summer it's closer it's near towards the sun.
5	B	I think I get it. Can you just draw a picture so I'm completely sure?
6	A	Okay. The sun's in the middle and uh-
7	B	Mmhm. Nice sun.

- 8 A and then the uh the Earth kind of orbits around it
- 9 B uh huh
- 10 A And um like say at one it's probably more of an ovally type thing -
- 11 B Mmhm
- 12 A In the winter, and uh er probably this will be winter *((moves pen tip to the opposite side of the orbit and draws a new Earth))* since it's further away
- 13 B Mmhm
- 14 A See, that's, winter would be like, the Earth orbits around the sun. Like summer is the closest to the sun. Spring is kind of a little further away, and then like fall is further away than spring, but like not as far as winter
- 15 B Mm hmm
- 16 A and then winter is the furthest.
- 17 B mm hmm
- 18 A So the sun doesn't, it like the flashlight and the bulb *((hand opening gesture over the sun, as if her fingers were the sun's rays spreading out))*, it hits summer,
- 19 B Mm hmm
- 20 A the lines like fade in *((draws fading lines from sun to summer))*, and get there closer, like quicker
- 21 B mm hmm
- 22 A And by the time they get there *[winter]*, they kinda fade and it's gets a lot colder for winter

- 23 B mm hmm
- 24 A And spring it's uh kinda *((gesturing between the sun and the earth labeled spring))* between the two *[between winter and summer]* and same for fall
- 25 B Mm hmm. Mm hmm. Um, Is this something - have you done this already for your class – is that you know this from?
- 26 A Uh, kind of, like from first and second grade I remember the time that the Earth orbiting and whatnot.
- 27 B mm hmm, mm hmm. Okay. So that makes a lot of sense. Um. One thing I wanted to ask you though about though was one thing that you might have heard is that at the same time - and you can tell me if you've heard this - when it's summer here *((B taps the table top))*, it's actually winter in Australia.
- 28 A mm hmm
- 29 B Have you heard that before?
- 30 A Yeah.
- 31 B So I was wondering if your picture the way you drew it can explain that or if that's a problem for your picture.
- 32 A Uhhhh. Idea. I need another picture.
- 33 B Okay. So is that a problem for your picture?
- 34 A Yeah, that is. Um, ok. *((A draws in a new sun, with smiley face, on her new piece of paper.))* There is like the sun. And okay. Yeah. *((A drawing a new elliptical orbit around the sun.))* I remember that now cause, um, it's like, as the world is

rotating, or as it's orbiting

35 B Mm hmm

36 A it's rotating too. So uh, I don't really – I guess I don't really understand it. Um.

37 B Well, you're saying as the Earth is going around here *((B sweeps once around the orbit A has drawn.))* it's doing what?

38 A It's like spinning. *((A again makes the quick "rotating" gesture between her thumb and first finger and she traces out the drawn orbit.))* Because it's. That's how it's day and night too.

39 B I see. It's spinning like a top. *((B makes a "spinning" gesture above A's diagram.))*

40 A Yeah.

41 B Okay.

42 A So, I guess I really don't understand it, that much. But. Uum. Yeah, I have heard that *[that when it is summer in Chicago, it is winter in Australia]*, cause I was supposed to go to Australia this summer

43 B Uh huh.

44 A but it was going to be winter

45 B Uh huh.

46 A when I was going, but uh, their winters are really warm. So,

47 B Mm hmm. So you're thinking that somehow the spinning, you thought that somehow if you take into account the fact that the Earth is also spinning, that

might help to explain why it's summer and winter at different times

48 A Uh - yeah.

49 B That's what you were thinking?

50 A Uh, kinda. Yeah.

51 B Just to be clear, what was – What was the problem with *this* picture for the-

52 A Because, yeah I rethought that and it looks really stupid to me because um  
summer is really close but then how could it be like winter on the other side.

Well. How could it be winter on the other side if it's really close here (*pointing to summer earth*), and how could it be really warm if this (*pointing to winter earth*) is this far away. I don't know. That looks really dumb to me now. So.

53 B It doesn't look really dumb to me. A lot of people explain it this way. Um. Okay, I'm not going to give away answers yet. You can find this out – you can find this out in your class.

### Rules for detecting contradictions

The system uses the following pairs of statement patterns to detect contradictions. We do not believe this list is complete for all tasks, but it is complete for the tasks involved in the simulations in Chapters 5-8. Each symbol beginning with a question mark (?) is a variable.

Statement 1	Statement 2
?x	(not ?x)
(greaterThan ?x ?y)	(lessThanOrEqualTo ?x ?y)

---

<code>(lessThan ?x ?y)</code>	<code>(greaterThanOrEqualTo ?x ?y)</code>
<code>(greaterThan ?x ?y)</code>	<code>(greaterThan ?y ?x)</code>
<code>(lessThan ?x ?y)</code>	<code>(lessThan ?y ?x)</code>

---

Note that there are rules for inferring `lessThanOrEqualTo` from `lessThan` and `equalTo`, and likewise for `greaterThanOrEqualTo`. Also, contradictory quantity changes are covered by the above ordinal relation pairs, since a quantity  $q$ 's continuous change in value is represented as an ordinal relation describing its derivative. For example, if  $q$  is increasing, we represent this as `(greaterThan (DerivativeFn q) 0)`. This means that if the system infers that a quantity is increasing and decreasing in the same time interval (e.g., in the seasons simulation in Chapter 6), it can detect the contradiction with the above rules.

### **Sentences from a textbook passage about the circulatory system**

These sentences were used to generate the instructional knowledge for the simulation in Chapter 7. Sentence numbers correspond to the sentence numbers in Chi et al. (2001). These sentences comprise the “structure” portion of the passage (Chi et al., 1994a).

1. The septum divides the heart lengthwise into two sides.
2. The right side pumps blood to the lungs, and the left side pumps blood to the other parts of the body.
3. Each side of the heart is divided into an upper and a lower chamber.
4. Each lower chamber is called a ventricle.
5. In each side of the heart blood flows from the atrium to the ventricle.

6. One-way valves separate these chambers and prevent blood from moving in the wrong direction.
7. The atrioventricular valves (a-v) separate the atria from the ventricles.
8. The a-v valve on the right side is the tricuspid valve, and the a-v valve on the left is the bicuspid valve.
9. Blood also flows out of the ventricles.
10. Two semilunar (s-l) valves separate the ventricles from the large vessels through which blood flows out of the heart.
11. Each of the valves consists of flaps of tissue that open as blood is pumped out of the ventricles.
12. Blood returning to the heart, which has a high concentration, or density, of carbon dioxide and a low concentration of oxygen, enters the right atrium.
13. The atrium pumps it through the tricuspid valve into the right ventricle.
14. The muscles of the right ventricle contract and force the blood through the right semilunar valve and into vessels leading to the lungs.
15. Each upper chamber is called an atrium.
16. In the lungs, carbon dioxide leaves the circulating blood and oxygen enters it.
17. The oxygenated blood returns to the left atrium of the heart.
18. The oxygenated blood is then pumped through the bicuspid valve into the left ventricle.
19. Strong contractions of the muscles of the left ventricle force the blood through the semilunar valve, into a large blood vessel, and then throughout the body.